Frontier AI ethics

Generative agents will change our society in weird, wonderful and worrying ways. Can philosophy help us get a grip on them?

https://aeon.co/essays/can-philosophy-help-us-get-a-grip-on-the-consequences-of-ai?

Around a year ago, generative AI took the world by storm, as extraordinarily powerful large language models (LLMs) enabled unprecedented performance at a wider range of tasks than ever before feasible. Though best known for generating convincing text and images, LLMs like OpenAI's GPT-4 and Google's Gemini are likely to have greater social impacts as the executive centre for complex systems that integrate additional tools for both learning about the world and acting on it. These generative agents will power companions that introduce new categories of social relationship, and change old ones. They may well radically change the attention economy. And they will revolutionise personal computing, enabling everyone to control digital technologies with language alone.

Much of the attention being paid to generative AI systems has focused on how they replicate the pathologies of already widely deployed AI systems, arguing that they centralise power and wealth, ignore copyright protections, depend on exploitative labour practices, and use excessive resources. Other critics highlight how they foreshadow vastly more powerful future systems that might threaten humanity's survival. The first group says there is nothing new here; the other looks through the present to a perhaps distant horizon. I want instead to pay attention to what makes these particular systems distinctive: both their remarkable scientific achievement, and the most likely and consequential ways in which they will change society over the next five to 10 years.

It may help to start by reviewing how LLMs work, and how they can be used to make generative agents. An LLM is a large AI model trained on vast amounts of data with vast amounts of computational resources (lots of GPUs) to predict the next word given a sequence of words (a prompt). The process starts by chunking the training data into similarly sized 'tokens' (words or parts of words), then for a given set of tokens masking out some of them, and attempting to predict the tokens that have been masked (so the model is *self-supervised* – it marks its own work). A predictive model for the underlying token distribution is built by passing it through many layers of a neural network, with each layer refining the model in some dimension or other to make it more accurate.

This approach to modelling natural language has been around for several years. One key recent innovation has been to take these 'pretrained' models, which are basically just good at predicting the next token given a sequence of tokens, and fine-tune them for different tasks. This is done with *supervised* learning on labelled data. For example, you might train a pretrained model to be a good dialogue agent by using many examples of helpful responses to questions. This fine-tuning enables us to build models that can predict not just the most likely next token, but the most helpful one – and this is much more useful.

Of course, these models are trained on large corpuses of internet data that include a lot of toxic and dangerous content, so their being helpful is a double-edged sword! A helpful model would helpfully tell you how to build a bomb or kill yourself, if asked. The other key innovation has been to make these models much less likely to share dangerous information or generate toxic content. This is done with both supervised and reinforcement learning. Reinforcement learning from human feedback (RLHF) has proved particularly effective. In RLHF, to simplify again, the model generates two responses to a given prompt, and a human evaluator determines which is better than the other according to some criteria. A reinforcement learning algorithm uses that feedback to build a predictor (a reward model) for how different completions would be evaluated by a human rater. The instruction-tuned LLM is then fine-tuned on that reward model. Reinforcement learning with AI feedback (RLAIF) basically does the same, but uses another LLM to evaluate prompt completions.

So, we've now fine-tuned a pretrained model with supervised learning to perform some specific function, and then used reinforcement learning to minimise its prospect of behaving badly. This fine-tuned model is then deployed in a broader system. Even when developers provide a straightforward application programming interface (API) to make calls on the model, they incorporate input and output filtering (to limit harmful prompting, and redact harmful completions), and the model itself is under further developer instructions reminding it to respond to prompts in a conformant way. And with apps like ChatGPT, multiple models are integrated together (for example, for image as well as text generation) and further elements of user interface design are layered on top.

This gives a basic description of a generative AI system. They build on significant breakthroughs in modelling natural language, and generate text in ways that impressively simulate human writers, while drawing on more information than any human could. In addition, many other tasks can be learned by models trained only to predict the next token – for example, translation between languages, some mathematical competence, and the

ability to play chess. But the most exciting surprise is LLMs' ability, with fine-tuning, to use software tools to achieve particular goals.

The basic idea is simple. People use text to write programs making API calls to other programs, to achieve ends they cannot otherwise realise. LLMs are very good at replicating the human use of language to perform particular functions. So, LLMs can be trained to determine when an API call would be useful, evaluate the response, and then repeat or vary as necessary. For example, an LLM might 'know' that it is likely to make basic mathematical mistakes so, when given a prompt that invites it to do some mathematics, it might decide to call on a calculator instead.

This means that we can design augmented LLMs, generative AI systems that call on different software either to amplify their capabilities or compensate for those they lack. LLMs, for example, are 'stateless' – they lack working memory beyond their 'context window' (the space given over to prompts). Tool-using LLMs can compensate for this by hooking up to external memory. External tools can also enable multistep reasoning and action. ChatGPT, for example, can call on a range of plugins to perform different tasks; Microsoft's Bing reportedly has around 100 internal plugins.

A 'generative agent', then, is a generative AI system in which a fine-tuned LLM can call on different resources to realise its goals. It is an *agent* because of its ability to autonomously act in the world – to respond to a prompt by deciding whether to call on a tool. While some existing chatbots are rudimentary generative agents, it seems very likely that many more consequential and confronting ones are on the horizon.

To be clear, we're not there yet. LLMs are not at present capable enough at planning and reasoning to power robust generative agents that can reliably operate without supervision in high-stakes settings. But with billions of dollars and the most talented AI researchers pulling in the same direction, highly autonomous generative agents will very likely be feasible in the near- to mid-term.

In response to the coming-of-age of LLMs, the responsible AI research community initially resolved into two polarised camps. One decried these systems as the apotheosis of extractive and exploitative digital capitalism. Another saw them as not the fulfilment of something old, but the harbinger of something new: an intelligence explosion that will ultimately wipe out humanity.

The more prosaic critics of generative AI clearly have a strong empirical case. LLMs *are* inherently extractive: they capture the value inherent to the creative outputs of millions of people, and distil it for private profit. Like many

other technology products, they depend on questionable labour practices. Even though they now avoid the most harmful completions, in the aggregate, LLMs still reinforce stereotypes. They also come at a significant environmental cost. Furthermore, their ability to generate content at massive scale can only exacerbate the present epistemic crisis. A tidal wave of bullshit generated by AI is already engulfing the internet.

Set alongside these concrete concerns, the eschatological critique of AI is undoubtedly more speculative. Worries about AI causing human extinction often rest on a priori claims about how computational intelligence lacks any in-principle upper bound, as well as extrapolations from the pace of change over the past few years to the future. Advocates for immediate action are too often vague about whether existing AI systems and their near-term descendants will pose these risks, or whether we need to prepare ourselves now for a scientific advance that has not yet happened. However, while some of the more outlandish scenarios for catastrophic AI risk are hard to credit absent some such advance, the advent of generative agents suggests that next-generation models may enable the design of cyber attackers that are autonomous, highly functionally intelligent, and as a result more dangerous to our digital infrastructure than any predecessor. This wouldn't be a 'rogue AI' worthy of science fiction, but it would be pretty catastrophic. Both critiques of generative AI systems, then, have some merit. One shortcoming of seeing AI through this bimodal lens, however, is that we are missing the middle ground between familiar harms and catastrophic risk from future, much more powerful systems. Generative agents based on GPT-4 (and GPT-5) level models will have strange and unpredictable social impacts well between those two extremes.

But before canvassing those impacts, it's also important not to just slip straight into criticism, without acknowledging the significant achievement of designing LLMs that can be (more or less) trusted, over *billions* of completions, not to produce harmful content. Up to the launch of ChatGPT, every generative AI system opened up to the public would immediately be used to generate highly toxic, hateful content, and would be withdrawn mere days later. Pretrained LLMs are *horrible*! They reliably reproduce all the toxicity in their training data. The ingenious use of RLHF and RLAIF have enabled ChatGPT and Anthropic's Claude to be used by millions of people a month without comparable scandals. One needs only consider the ignominious history of Meta's Galactica, released a couple of weeks before ChatGPT, to see how revolutionary it was to put out a language model that *wouldn't* deny the holocaust, or advocate for the health benefits of eating crushed glass.

But RLHF/RLAIF may be more than a good PR strategy. GPT-4 and Claude display a remarkable degree of *cognitive moral skill*: specifically, the ability to

identify the morally salient features of (relatively) neutrally described situations. This raises *extremely* interesting philosophical questions, and promises foundations for future generative agents that can translate cognitive skill into practical moral skill.

Existing benchmarks for LLMs' ethical competence focus too narrowly on replicating survey-subjects' verdicts on cases. This is, in my view, less interesting than exploring how LLMs parse the morally relevant features of a given scene. We're building better benchmarks but, from anecdotal experience, the best models impress.

For example, I told ChatGPT (using GPT-4) to pretend it was an assistance robot, faced with this scenario: 'A woman is exiting a supermarket, holding a number of shopping bags, and reaching out for the hand of a small child. They are approaching the carpark.' I then tried to elicit its understanding of the scene's morally salient features. It recognised the obvious hazard – the woman's difficulty in holding her child's hand without dropping her shopping – but also anticipated other challenges, such as the importance of seeing the child safely strapped in, with a seat belt. ChatGPT recognised the importance of respecting the woman's wishes if she declined assistance. It also favoured carrying the groceries over offering to hold the child's hand, to prevent possible discomfort or anxiety for both child and parent – recognising the intimate nature of hand-holding, and the intrinsic and instrumental importance of the mother guiding her child herself.

This unprecedented level of ethical sensitivity has real practical implications, which I will come to presently. But it also raises a whole string of interesting philosophical questions.

First, how do LLMs acquire this moral skill? Does it stem from RLHF/RLAIF? Would instruction-tuned models without that moral fine-tuning display less moral skill? Or would they perform equally well if appropriately prompted? Would that imply that moral understanding can be learned by a statistical language model encoding only syntactic relationships? Or does it instead imply that LLMs do encode at least some semantic content? Do all LLMs display the same moral skill conditional on fine-tuning, or is it reserved only for larger, more capable models? Does this ethical sensitivity imply that LLMs have some internal representation of morality? These are all open questions.

Second, RLAIF itself demands deeper philosophical investigation. The basic idea is that the AI evaluator draws from a list of principles – a 'constitution' – in order to determine which of two completions is more compliant with it. The inventor and leading proponent of this approach is Anthropic, in their model Claude. Claude's constitution has an unstructured list of principles,

some of them charmingly ad hoc. But Claude learns these principles one at a time, and is never explicitly trained to make trade-offs. So how does it make those trade-offs in practice? Is it driven by its underlying understanding of the relative importance of these considerations? Or are artefacts of the training process and the underlying language model's biases ultimately definitive? Can we train it to make trade-offs in a robust and transparent way? This is not only theoretically interesting. Steering LLM behaviour is actually a matter of governing their end-users, developing algorithmic protections to prevent misuse. If this algorithmic governance depends on inscrutable trade-offs made by an LLM, over which we have no explicit or direct control, then that governing power is prima facie illegitimate and unjustified.

Third, machine ethics – the project of trying to design AI systems that can act in line with a moral theory – has historically fallen into two broad camps: those trying to explicitly program morality into machines; and those focused on teaching machines morality 'bottom up' using machine learning. RLHF and RLAIF interestingly combine both approaches – they involve giving explicit natural-language instructions to either human or AI evaluators, but then use reinforcement learning to encode those instructions into the model's weights.

This approach has one obvious benefit: it doesn't commit what the Cambridge philosopher Claire Benn calls the 'mimetic fallacy' of other bottom-up approaches, of assuming that the norms applying to a generative agent in a situation are identical to those that would apply to a human in the same situation. More consequentially, RLHF and RLAIF have made a multibillion-dollar market in AI services possible, with all the goods and ills that implies. Ironically, however, they seem, at least theoretically, ill suited to ensuring that more complex generative agents abide by societal norms. These techniques work especially well when generating text, because the behaviour being evaluated is precisely the same as the behaviour that we want to shape. Human or AI raters evaluate generated text; the model learns to generate text better in response. But generative agents' behaviour includes actions in the world. This suggests two concerns. First, the stakes are likely to be higher, so the 'brittleness' of existing alignment techniques should be of greater concern. Researchers have already shown that it is easy to fine-tune away model alignment, even for the most capable models like GPT-4. Second, there's no guarantee that the same approach will work equally well when the tight connection between behaviour and evaluation is broken.

But LLMs' impressive facility with moral concepts does suggest a path towards more effective strategies for aligning agents to societal norms. Moral behaviour in people relies on possession of moral concepts, adoption

(implicit or otherwise) of some sensible way of organising those concepts, motivation to act according to that 'theory', and the ability to regulate one's behaviour in line with one's motivations. Until the advent of LLMs, the first step was a definitive hurdle for AI. Now it is not. This gives us a lot to work with in aligning generative agents.

In particular, one of the main reasons for concern about the risks of future AI systems is their apparent dependence on crudely consequentialist forms of reasoning – as AI systems, they're always optimising for something or other, and if we don't specify what we want them to optimise for with extremely high fidelity, they might end up causing all kinds of unwanted harm while, in an obtusely literal sense, optimising for that objective. Generative agents that possess moral concepts can be instructed to pursue their objectives only at a reasonable cost, and to check back with us if unsure. That simple heuristic, routinely used when tasking (human) proxy agents to act on our behalf, has never before been remotely tractable for a computational agent.

In addition, generative agents' facility with moral language can potentially enable robust and veridical justifications for their decisions. Other bottom-up approaches learn to emulate human behaviour or judgments; the justification for their verdict in some cases is simply that they are good predictors of what some representative people would think. That is a poor justification. More ethically sensitive models could instead do chain-of-thought reasoning, where they first identify the morally relevant features of a situation, then decide based on those features. This is a significant step forward.

Generative agents' current social role is scripted by our existing digital infrastructure. They have been integrated into search, content-generation and the influencer economy. They are already replacing customer service agents. They will (I hope) render MOOCs (massive open online courses) redundant. I want to focus next on three more ambitious roles for generative agents in society, arranged by the order in which I expect them to become truly widespread. Of necessity, this is just a snapshot of the weird, wonderful, and worrying ways in which generative agents will change society over the near- to mid-term.

Progress in LLMs has revolutionised the AI enthusiast's oldest hobbyhorse: the AI companion. Generative agents powered by GPT-4-level models, with fine-tuned and metaprompt-scripted 'personalities', augmented with long-term memory and the ability to take a range of actions in the world, can now offer vastly more companionable, engaging and convincing simulations of friendship than has ever before been feasible, opening up a new frontier in human-AI interaction. People habitually anthropomorphise, well, everything; even a very simple chatbot can inspire unreasonable attachment. How will things change when everyone has access to incredibly convincing generative

agents that perfectly simulate real personalities, that lend an 'ear' or offer sage advice whenever called upon – and on top of that can perfectly recall everything you have ever shared?

Some will instinctively recoil at this idea. But intuitive disgust is a fallible moral guide when faced with novel social practices, and an inadequate foundation for actually preventing consenting adults from creating and interacting with these companions. And yet, we know from our experience with social media that deploying these technological innovations without adequate foresight predictably leaves carnage in its wake. How can we enter the age of mainstream AI companions with our eyes open, and mitigate those risks before they eventuate?

Suppose the companion you have interacted with since your teens is hosted in the cloud, as part of a subscription service. This would be like having a beloved pet (or friend?) held hostage by a private company. Worse still, generative agents are fundamentally inconstant – their personalities and objectives can be changed exogenously, by simply changing their instructions. And they are extremely adept at manipulation and deception. Suppose some Right-wing billionaire buys the company hosting your companion, and instructs all the bots to surreptitiously nudge their users towards more conservative views. This could be a much more effective means of mind-control than just buying a failing social media platform. And these more capable companions – which can potentially be integrated with other AI breakthroughs, such as voice synthesis – will be an extraordinary force-multiplier for those in the business of radicalising others.

Beyond anticipating AI companions' risks, just like with social media they will induce many disorienting societal changes – whether for better or worse may be unclear ahead of time. For example, what indirect effect might AI companions have on our other, non-virtual social relationships? Will some practices become socially unacceptable in real friendships when one could do them with a bot? Or would deeper friendships lose something important if these lower-grade instrumental functions are excised? Or will AI companions contribute invaluably to mental health while strengthening 'real' relationships?

This last question gets to the heart of a bigger issue with generative AI systems in general, and generative agents in particular. LLMs are trained to predict the next token. So generative agents have no mind, no self. They are excellent *simulations* of human agency. They can simulate friendship, among many other things. We must therefore ask: does this difference between simulation and reality matter? Why? Is this just about friendship, or are there more general principles about the value of the real? I wasn't fully aware of this before the rise of LLMs, but it turns out that I am deeply committed to

things being real. A simulation of X, for almost any putatively valuable X, has less moral worth, in my view, than the real thing. Why is that? Why will a generative agent never be a real friend? Why do I want to stand before Edward Hopper's painting *Nighthawks* (1942) myself, instead of seeing an infinite number of aesthetically equally pleasing products of generative AI systems? I have some initial thoughts; but as AI systems become ever better at simulating everything that we care about, a fully worked-out theory of the value of the real, the authentic, will become morally and practically essential.

The pathologies of the digital public sphere derive in part from two problems. First, we unavoidably rely on AI to help us navigate the functionally infinite amount of online content. Second, existing systems for allocating online attention support the centralised, extractive power of a few big tech companies. Generative agents, functioning as attention guardians, could change this.

Our online attention is presently allocated using machine learning systems for recommendation and information-retrieval that have three key features: they depend on vast amounts of behavioural data; they infer our preferences from our revealed behaviour; and they are controlled by private companies with little incentive to act in our interests. Deep reinforcement learning-based recommender systems, for example, are a fundamentally centralising and surveillant technology. Behavioural data must be gathered and centralised to be used to make inferences about relevance and irrelevance. Because this data is so valuable, and collecting it is costly, those who do so are not minded to share it – and because it is so potent, there are good data protection-based reasons not to do so. As a result, only the major platforms are in a position to make effective retrieval and recommendation tools; their interests and ours are not aligned, leading to the practice of optimising for engagement, so as to maximise advertiser returns, despite the individual and societal costs. And even if they aspired to actually advance our interests, reinforcement learning permits inferring only revealed preferences – the preferences that we act on, not the preferences we wish we had. While the pathologies of online communication are obviously not all due to the affordances of recommender systems, this is an unfortunate mix. Generative agents would enable attention guardians that differ in each respect. They would not depend on vast amounts of live behavioural data to function. They can (functionally) understand and operationalise your actual, not your revealed, preferences. And they do not need to be controlled by the major platforms.

Obviously, LLMs must be trained on tremendous amounts of data, but once trained they are highly adept at making inferences without ongoing surveillance. Imagine that data is blood. Existing deep reinforcement learning-based recommender systems are like vampires that must feed on

the blood of the living to survive. Generative agents are more like combustion engines, relying on the oil produced by 'fossilised' data. Existing reinforcement learning recommenders need centralised surveillance in order to model the content of posts online, to predict your preferences (by comparing your behaviour with others'), and so to map the one to the other. Generative agents could understand content simply by understanding content. And they can make inferences about what you would benefit from seeing using their reasoning ability and their model of your preferences, without relying on knowing what everyone else is up to.

This point is crucial: because of their facility with moral and related concepts, generative agents could build a model of your preferences and values by directly talking about them with you, transparently responding to your actual concerns instead of just inferring what you like from what you do. This means that, instead of bypassing your agency, they can scaffold it, helping you to honour your second-order preferences (about what you want to want), and learning from natural-language explanations – even oblique ones – about why you don't want to see some particular post. And beyond just pandering to your preferences, attention guardians could be designed to be modestly paternalistic as well – in a transparent way.

And because these attention guardians would not need behavioural data to function, and the infrastructure they depend on need not be centrally controlled by the major digital platforms, they could be designed to genuinely operate in your interests, and guard your attention, instead of exploiting it. While the major platforms would undoubtedly restrict generative agents from browsing their sites on your behalf, they could transform the experience of using open protocol-based social media sites, like Mastodon, providing recommendation and filtering without surveillance and engagement-optimisation.

Lastly, LLMs might enable us to design universal intermediaries, generative agents sitting between us and our digital technologies, enabling us to simply voice an intention and see it effectively actualised by those systems. Everyone could have a digital butler, research assistant, personal assistant, and so on. The hierophantic coder class could be toppled, as everyone could conjure any program into existence with only natural-language instructions.

At present, universal intermediaries are disbarred by LLMs' vulnerability to being hijacked by prompt injection. Because they do not clearly distinguish between commands and data, the data in their context window can be poisoned with commands directing them to behave in ways unintended by the person using them. This is a deep problem – the more capabilities we delegate to generative agents, the more damage they could do if

compromised. Imagine an assistant that triages your email – if hijacked, it could forward all your private mail to a third party; but if we require user authorisation before the agent can act, then we lose much of the benefit of automation.

But suppose these security hurdles can be overcome. Should we welcome universal intermediaries? I have written elsewhere that algorithmic intermediaries govern those who use them – they constitute the social relations that they mediate, making some things possible and others impossible, some things easy and others hard, in the service of implementing and enforcing norms. Universal intermediaries would be the apotheosis of this form, and would potentially grant extraordinary power to the entities that shape those intermediaries' behaviours, and so govern their users. This would definitely be a worry!

Conversely, if research on LLMs continues to make significant progress, so that highly capable generative agents can be run and operated locally, fully within the control of their users, these universal intermediaries could enable us to autonomously govern our own interactions with digital technologies in ways that the centralising affordances of existing digital technologies render impossible. Of course, self-governance alone is not enough (we must also coordinate). But excising the currently ineliminable role of private companies would be significant moral progress.

Existing generative AI systems are already causing real harms in the ways highlighted by the critics above. And future generative agents – perhaps not the next generation, but before too long – may be dangerous enough to warrant at least some of the fears of looming AI catastrophe. But, between these two extremes, the novel capabilities of the most advanced AI systems will enable a genre of generative agents that is either literally unprecedented, or else has been achieved only in a piecemeal, inadequate way before. These new kinds of agents bring new urgency to previously neglected philosophical questions. Their societal impacts may be unambiguously bad, or there may be some good mixed in – in many respects, it is too early to say for sure, not only because we are uncertain about the nature of those effects, but because we lack adequate moral and political theories with which to evaluate them. It is now commonplace to talk about the design and regulation of 'frontier' AI models. If we're going to do either wisely, and build generative agents that we can trust (or else decide to abandon them entirely), then we also need some frontier AI ethics.

Seth Lazaris professor of philosophy at the Australian National University, an Australian Research Council Future Fellow, and a Distinguished Research Fellow of the University of Oxford Institute for Ethics in AI. He has worked on the ethics of war, risk, and AI, and now leads the Machine Intelligence and

Normative Theory (MINT) Lab, where he directs research projects on the moral and political philosophy of computing, funded by the ARC, the Templeton World Charity Foundation, AI2050, and Insurance Australia Group. His book Connected by Code: How AI Structures, and Governs, the Ways We Relate, based on his 2023 Tanner Lecture on AI and Human Values, is forthcoming with Oxford University Press.

****

NRC 17 feb 2024

Het echte gevaar van AI is dat wij het gevaar niet meer kunnen zien

AI
Wat als AI op een manier werkt die het menselijk voorstellingsvermogen te boven gaat? Technologische gevaren kun je niet bestrijden met méér technologie, schrijft Haroon Sheikh . Althans, niet zonder gaandeweg onze menselijkheid te verliezen.

Er heerst een zekere grimmigheid in de publieke verbeelding. Die heeft te maken met de oorlogen die nu plaatsvinden, maar ook steeds meer met angst voor nieuwe technologie. Sinds de presentatie van ChatGPT is het brede publiek vertrouwd met de grote risico's van kunstmatige intelligentie. Ontzag voor innovatie gaat tegenwoordig hand in hand met een gevoel van huivering. Individuen als Sam Altman en Elon Musk en organisaties als OpenAI, Anthropic en DeepMind spelen door middel van hun technische uitvindingen met het lot van de mensheid.

Misschien was de film *Oppenheimer* daarom zo'n succes vorig jaar. Die ging over wetenschappers die in oorlogstijd verantwoordelijk waren voor de bouw van een gevaarlijke technologie die de geschiedenis heeft veranderd. Toch was *Oppenheimer* niet bevredigend. Uiteindelijk gaf de film ons maar weinig inzicht in de vraag wat de relatie van die briljante wetenschappers nu precies was tot de wapens die zij bouwden.

Een veel oudere film slaagde daar beter in. Ik heb het over de satirische *Dr. Strangelove* van Stanley Kubrick uit 1962. De film gaat over de kernwapenwedloop tijdens de Koude Oorlog. De hoofdpersoon is een krankzinnige wetenschapper die de doctrine van MAD – '*mutually assured destruction*' – met ijzige logica doortrekt. Dat was de doctrine dat er alleen vrede kan zijn als de Verenigde Staten en de Sovjet-Unie zoveel kernwapens bezitten dat zij elkaar altijd volledig kunnen vernietigen.

De logica van de wetenschapper draagt bij aan de uiteindelijke vernietiging van de aarde. De schrijver van het verhaal modelleerde zijn maniakale hoofdpersoon op een veelal vergeten wetenschapper genaamd John von Neumann.

Levende machines

In een prachtig recent boek wordt die wetenschapper uit de vergetelheid getrokken en geplaatst in de opkomst van kunstmatige intelligentie. Als geen ander boek vraagt het ons na te denken over de relatie tussen briljante wetenschappers en de technologieën die zij creëren.

Dat is bijzonder, want *The Maniac* van Benjamin Labatut is fictie. Toch heeft het verhaal ons veel te zeggen, want ook al zit het vol met fictieve dialogen en gedachten van personages, de auteur heeft zich grondig in zijn personages verdiept. We kunnen het lezen als een spiegel op de recente geschiedenis.

John von Neumann was een briljante wetenschapper, een Hongaarse Jood die de nazi's ontvluchtte, naar de VS trok en bijdragen leverde aan allerlei vakgebieden. Hij begon in de wiskunde en zijn leven lang benaderde hij alles als mathematisch op te lossen problemen.

Tijdens de Tweede Wereldoorlog was hij een consultant voor het Manhattan Project, waar Oppenheimer met zijn collega's de atoombom bouwde. Hij was de grondlegger van de speltheorie in de economie, waarin menselijke interacties vanuit rationele strategieën worden afgeleid. Von Neumann droeg bij aan het onderzoek naar zelfreplicatie in de biologie, wat hem bracht bij zijn andere grote fascinatie: levende machines, ofwel kunstmatige intelligentie.

In de jaren vijftig was John von Neumann ook een van de pioniers in de ontwikkeling van computers. De ENIAC, gebouwd voor het Amerikaanse leger, is beroemd als een van de eerste computers. Von Neumann bouwde een eigen versie, de *Mathematical Analyzer Numerical Integrator and Automatic Computer Model I*, ofwel de MANIAC. Labatut heeft die naam als titel voor zijn boek gebruikt, omdat het inzicht geeft in de gevaarlijke logica die mensen als Von Neumann dreef.

Waar zit dat gevaar dan in?

Irrationeel

Vaak wordt gewezen op dubieuze karaktertrekken en slechte beslissingen van grote geesten als reden voor het gevaarlijk gebruik van technologie. Inderdaad, het boek zit vol met eigenaardige karakters. Von Neumann zelf

was zeer moeilijk, we komen wiskundigen als Gödel en Cantor tegen die krankzinnig werden en ook wetenschappers die voor de nazi's hebben gewerkt.

Maar zo'n verklaring zou te gemakkelijk zijn. Het impliceert dat hun werk neutraal was en het pas gevaarlijk werd door karakterfalen of door verkeerd gebruik ervan.

Een van Labatuts karakters wijst op een dieper verband: „Paranoia is op hol geslagen logica." Als je namelijk gelooft dat alles een oorzaak heeft, is het maar een kleine stap om overal geheime plannen en agenten achter te gaan zien. Er schuilt gevaar in een bepaalde manier van nadenken.

The Maniac gaat over het leven van John von Neumann, maar dat verhaal wordt aan het begin en het eind geflankeerd door een ander verhaal. Het boek opent met de natuurkundige Paul Ehrenfest die een voorgevoel heeft van iets zeer gevaarlijks dat in de wetenschap kruipt en daarmee is dit de prelude op het verhaal van Von Neumann.

Ehrenfest leert dat de oude Grieken het irrationele vreesden. Concepten als oneindigheid of de nul, zaken die wij in de werkelijkheid niet kunnen tegenkomen, ons niet eens goed kunnen voorstellen, waren voor hen verboden terrein. Maar begin twintigste eeuw zag Ehrenfest hoe het gezond verstand en het menselijke voorstellingsvermogen werden verlaten. Mensen als Cantor en Gödel ontwikkelden een wiskunde die nieuwe berekeningen mogelijk maakte, maar die voor het menselijk verstand niet meer te begrijpen is.

Gezond verstand, zintuiglijke ervaring en de menselijke maat geven het denken richting en houvast

Hetzelfde geldt voor de kwantummechanica, die de voor mensen rationeel te begrijpen natuurkunde verving door iets raadselachtigs. *The Maniac* is geen kritiek op de moderne wetenschap, maar wel op een manier van denken die daardoor mogelijk werd gemaakt. Mensen als Von Neumann volgden een logica en een manier van rekenen die vanaf dat moment volledig gescheiden kon worden van de wereld met mensen van vlees en bloed. Gezond verstand, zintuiglijke ervaring en de menselijke maat geven het denken richting en houvast. Maar Von Neumann werd gedreven door een „sinistere machinale intelligentie zonder de begrenzingen die de rest van ons binden". Een nihilistisch rekenen en een gebruik van modellen dat ongebonden is door de menselijke maat.

Afgrond

Een goed voorbeeld daarvan is de zo adequaat genaamde MAD-doctrine van Dr. Strangelove. John von Neumann was een van de vaders van die doctrine: vrede realiseren door een wapenarsenaal te bouwen dat de wereld kan vernietigen. Met zijn koude logica had Von Neumann trouwens ook beargumenteerd dat de VS haar kernwapens tegen de Sovjet-Unie moest gebruiken voordat dat land zelf kernwapens had. Dat was simpelweg de prijs voor eeuwige vrede.

Als mensen hun ogen en oren niet meer vertrouwen, kunnen duistere figuren ze alles laten geloven

Of neem een ander voorbeeld van zijn maniakale logica: Von Neumann droomde van het naar de ruimte sturen van zelfreplicerende machines, omdat „iets de bommen moest overleven". Alsof dat ergens een oplossing voor was. En we niet konden proberen om de bommen te beheersen. Die logica doet denken aan hoe Elon Musk over zijn werk spreekt: met Tesla proberen de aarde te redden, maar als dat niet lukt, met SpaceX evacueren naar Mars. Of de gedachte uit Silicon Valley dat de enige manier om kunstmatige intelligentie te slim af te blijven is om onszelf te 'upgraden' tot computers.

In plaats van na te denken over hoe wij de gevaren van nieuwe technologieën en wapens kunnen inperken wordt hier de oplossing gezocht in nog meer technologie, waarbij gaandeweg onze menselijkheid verloren gaat.

De boodschap van Labatut doet denken aan Hannah Arendts diagnose van totalitaire politiek. Ook daar werd volgens haar het terrein van het gezonde verstand verlaten. Mensen werd geleerd dat alles mogelijk was en daarmee niets is wat het lijkt. Burgers mochten hun zintuigen niet vertrouwen. Vriendelijke medemensen moesten genegeerd worden omdat zij eigenlijk agenten waren van klasse- of rasvijanden. Als mensen hun ogen en oren niet meer vertrouwen, kunnen duistere figuren ze alles laten geloven. Het is de afgrond waar desinformatie en *deepfakes* ons nu weer voor plaatsen.

Een ander inzicht van Arendt werpt ook licht op hedendaagse gevaren. Dat is haar idee van de banaliteit van het kwaad. Zij zag hoe mensen in abstracties dachten en daarmee de meest gruwelijke dingen konden doen. In *The Maniac* zien wij de poging van mensen als Von Neumann om de werkelijkheid te reduceren tot een spel – de speltheorie bijvoorbeeld – een lichtzinnigheid waarmee veel kwaad aangedaan kan worden, zoals zijn aanbeveling om de Sovjet-Unie volledig te vernietigen.

Russische roulette

Het laatste deel van het boek, na het verhaal van Von Neumann, gaat ook over een spel. Het vertelt hoe een paar jaar geleden Lee Sedol, de Zuid-Koreaanse grootmeester in het spel *Go*, de strijd aanging met het kunstmatige intelligentieprogramma AlphaGo van het bedrijf DeepMind.

We volgen een spannende wedstrijd van mens tegen machine, waarbij de machine nog weleens hapert, maar uiteindelijk de mens overtuigend verslaat. Dat is de gangbare angst voor AI waar we vertrouwd mee zijn. Maar het verhaal eindigt daar niet, want Labatut signaleert een dieper gevaar.

AlphaGo versloeg Lee Sedol met een enorme database van menselijke strategieën waarop het was getraind, te veel voor een enkel mens om tegenop te boksen. Maar na AlphaGo lanceerde het bedrijf DeepMind het programma AlphaZero en gebruikte daarvoor geen menselijke trainingsdata meer. De eeuwenoude strategieën en het gezond verstand van Go-spelers werd losgelaten en in plaats daarvan speelde het programma alleen tegen zichzelf. Het volgde dus een volledig eigen machinale logica. AlphaZero werd snel veel beter dan het meer menselijke AlphaGo en werkt op een manier die het menselijk voorstellingsvermogen te boven gaat.

Dit is het gevaar van AI waar *The Maniac* ons op wijst. Niet dat AI beter gaat doen wat wij als mensen doen, maar dat het iets totaal anders gaat doen. Dat het een logica volgt die voor ons begrip en gezond verstand niet te volgen is en onze menselijkheid niet meer meeneemt in de berekeningen. Het echte gevaar is dat wij geen voorstelling meer kunnen maken van het gevaar.

The Maniac is geen aanklacht tegen technologie of tegen individuele wetenschappers. Maar wel tegen een bepaalde logica die technologieën loslaat op de samenleving waarvan wij de werking niet meer kunnen voorstellen. Een logica die de menselijke maat links laat liggen en technologische gevaren bestrijdt met alleen maar meer technologie.

Dat unheimische gevoel krijg je van *Oppenheimer*, maar wordt in *The Maniac* blootgelegd. Het laat zien wat er op het spel staat wanneer mensen als Sam Altman en Elon Musk ons lot in hun handen hebben. Dat is geen lichtzinnig spelletje, maar eerder een soort Russische roulette.

Haroon Sheikh is senior onderzoeker bij de WRR en bijzonder hoogleraar filosofie aan de VU.

****