Noreen Herzfeld

The Banality of Artificial Evil

Abstract

Can AI be aligned with human values? Hannah Arendt's examination of virtue in Nazi Germany suggests three stumbling blocks. First, Arendt argues that virtue is not rule-based. Arendt noted that social codes are insufficient as they can rapidly change and that particular cases require particular answers no general rules can predict. Instead, virtue relies on inner introspection, a dialogue with oneself that determines when one determines "I cannot do this." Such introspection requires a level of sentience and theory of mind that computers do not yet have. Finally, AI threatens the ultimate value, life itself, through its hidden usage of vast amounts of energy. As it is scaled up from a niche application to the general public, it will increasingly contribute to climate instability and thus to political and social instability.

1. Introduction: "Not My Problem"

At a conference on technology and faith a few years back, I was speaking with a fellow computer scientist who was enthusiastically describing her work programming robots to play soccer as a team. When I, as a Quaker, later spoke about my concerns regarding the development of lethal autonomous weapons, she wholeheartedly nodded in agreement. Later, when asked who was funding her research she replied, "The DoD (US Department of Defense), of course." When I asked, "The DoD wants robots that play soccer?", she looked at me blankly and then replied, "That's not my problem."

Seventy years ago, another person said the same thing: "It wasn't my problem. I was only doing my job." Hannah Arendt coined the

phrase "the banality of evil" as the subtitle of her groundbreaking examination of the trial of Otto Adolf Eichmann in Jerusalem. Eichmann presided over the transportation of millions of Jews to the concentration camps. Yet Arendt was astounded to find that while the deeds were monstrously lethal, "the doer […] was quite ordinary, commonplace, and neither demonic nor monstrous".[1] While Arendt notes that Eichmann was both evil and not well educated, she particularly critiques him as being thoughtless. Nor was Eichmann unique among his countrymen. Thousands of Germans participated in the Nazi death machine, most of them ordinary people doing their somewhat ordinary jobs.

Arendt noted "the phenomenon of evil deeds, committed on a gigantic scale, which could not be traced to any particularity of wickedness, pathology, or ideological conviction in the doer[s], only […] shallowness."[2] Shallowness could be used to describe far too much of AI technology today. Looking beyond the hype, we find AI beset by shallow algorithms, a shallow understanding of thinking, and a shallow consideration of what price we are paying, in terms of the environment, when we use AI as a tool or a diversion. There is no wickedness or pathology in AI, but there is a great deal of shallowness, a shallowness that we disregard at our peril.

2. Shallow values

The hype around AI rarely corresponds to its reality. Consider the algorithms that run social media. When Mark Zuckerberg developed Facebook, his stated goal was "to make the world more connected". Zuckerberg notes that he once thought "if we just gave people a voice and helped them connect, that would make the world better by itself".[3] Nor was he alone in thinking this. Theologian Ilia Delio still believes that the internet is bringing us closer to Teilhard de Chardin's vision of a humanity united in love and purpose: "Teilhard anticipated a new level of collective mind which he called the 'noosphere', from the Greek *nous* (mind). Computer technology has

1 *Arendt*, The Life of Mind, 4.
2 *Arendt,* Thinking and Moral Considerations, 417.
3 *Zuckerberg*, Bringing the World Closer Together.

initiated this next step of evolution [...] the natural culmination of evolution and not its termination."[4]

Well, not exactly united. Instead, we find social media sites populated by bots and running on algorithms that bring out far too many people's inner troll, inflame human emotions, divide us into hermetic social bubbles and propagate misinformation. The algorithms, which manipulate what we see and our emotional states, are hidden behind the screen; indeed, they are proprietary secrets. They are designed, first and foremost, with shallow goals—to keep us scrolling so that we will see more ads, each tailored to tempt us to buy products, bringing revenue to the advertisers and maintaining advertisers on the site. The goal is not to connect us but to disconnect us from our money. Dividing us into political camps or damaging the psyches of the young is not their primary intent. These are simply thoughtless byproducts.

Or consider generative AI. The rapid development of deep learning has led to recent advances in a variety of areas where AI seemed to have stalled. China, the US and the EU are pouring billions into AI research since, as a recent European Commission put it, "Like the steam engine or electricity in the past, AI is transforming our world, our society and our industry. Growth in computing power, availability of data and progress in algorithms have turned AI into one of the most strategic technologies of the 21st century."[5] Generative AI is expected to automate many white-collar jobs, boost corporate profits, solve intractable problems such as climate change, provide sociable care for the elderly, teach our children, revamp the process of producing poetry and art, and turn sexbots into romantic and chatty partners. Many, like Google engineer Blake Lemoine, see machine sentience right around the corner, if not already here.

That's the hype, anyway. As with social media, the reality is somewhat different. Programs like GPT-4 let you give a prompt, such as "describe Hannah Arendt's concept of the banality of evil". Scouring works on the internet, these programs put together text that is fairly indistinguishable from that of my undergraduate students. But, like Eichmann, these programs do not think critically, or, indeed, at all about what they are doing. Their design as language predictors

4 *Delio*, Re-Enchanting the Earth, xvii.

5 *European Commission*, Shaping Europe's Digital Future.

gives rise to convincing, human-like prose, yet they tend to "hallucinate", a polite term for bullshit. Without mental models of the world, they cannot distinguish between truth and falsehood, making them easily prompted to generate plausible misinformation.[6] They lack a moral compass. One chatbot suggested to a writer from the New York Times that he leave his wife, while another supported a Belgian man in committing suicide, hardly good advice for people whose situation the chatbot can neither fully comprehend nor contextualise.[7] In light of their capacity to generate misinformation as well as mess with our minds, over 30,000 AI developers, ethicists and concerned citizens world-wide (including luminaries such as Steve Wozniak, Elon Musk and Andrew Yang) have signed an open letter, which originated from the Future of Life Institute, calling for a moratorium on the further development of such programs to allow time for ethical safeguards to be erected. They ask that "AI research and development should be refocused on making today's powerful, state-of-the-art systems more accurate, safe, interpretable, transparent, robust, aligned, trustworthy, and loyal".[8]

Is this possible? What would it take to align these systems in such a way that we could deem them trustworthy, loyal or safe? How do we keep them from committing acts we generally consider evil? This is no simple task. Developers such as OpenAI have rushed to "erect guard rails" or patch up the most obvious problems, yet each patch seems to merely reveal another hole. That there is no simple code for virtue has long been known. In an address given at Riverside Church in Manhattan in 1966, at the height of the Vietnam War, Arendt noted the problem with such an approach:

> Particular questions must receive particular answers, and if the series of crises in which we have lived since the beginning of the century can teach us anything at all, it is, I think, the simple fact that there are no general standards to determine our judgements unfailingly, no general

6 Researchers from the Center for Countering Digital Hate, a UK-based nonprofit, found that Bard failed 78 of 100 test cases, generating plausible misinformation on a variety of subjects, including climate change, the war in Ukraine, vaccine efficacy and Black Lives Matter activists (see *Elliott*, It's Way Too Easy).

7 See *Walker*, Belgian Man Dies.

8 *Future of Life Institute*, Pause Giant AI Experiments: An Open Letter.

rules under which to subsume the particular cases with any degree of certainty.[9]

Arendt arrived at this conclusion through her examination of moral standards in Germany during the Nazi regime. She characterised their breakdown as originating in a lack of judgement and the concomitant abdication of personal responsibility.

This breakdown was not due to a lack of knowledge, nor was it the result of a lack of culture or artistic refinement. Arendt noted that the same person could spend an evening reading Goethe or listening to a Bach cantata and then send hundreds to the gas chambers the following morning. Standards of conduct previously thought to be "permanent and vital [...] and whose validity was supposed to be self-evident to every sane person" collapsed. This strengthened Arendt's propensity towards moral particularism. Indeed, she wondered whether virtue, understood as a set of moral precepts, is really nothing more than a set of customs, easily exchanged for another set at society's whim.[10]

When it came to participating in genocide, why did ordinary Germans like Eichmann not say "I cannot do this"? What motivated the few who refused to become a cog in the Nazi wheel, often at their own peril? According to Arendt, "they refused to murder, not so much because they still held fast to the command 'Thou shalt not kill', but because they were unwilling to live together with a murderer—themselves".[11] In other words, for Arendt, social norms or external codes cannot constitute reliable sources of virtue, for they can change, seemingly overnight. This experience of a rapid change in norms is not unique to Nazi Germany. In America, for example, only 30 percent of white evangelicals in 2011 agreed that "an elected official who commits an immoral act in their personal

9 *Arendt*, colloquium on The Crisis of Character of Modern Society.

10 Such a collapse is by no means unique to Nazi Germany. The horrors perpetrated on the citizens of Bucha and other Ukrainian towns have led observers to question whether there has been a similar breakdown in Russian culture and morality. Such breakdowns seem to be a common result of the dehumanisation of the other propounded in times of war.

11 *Arendt*, Responsibility and Judgement, 44.

life can still behave ethically and fulfil their duties in their public and professional life." A mere five years later, 72 percent agreed.[12]

Establishing "guard rails", adding programming that tells an AI to avoid certain words or subjects to implant norms and boundaries in generative AI, exhibits similar problems. First, such guard rails have shown themselves to be impervious to any sort of automation; hence, companies such as OpenAI or Google have resorted to armies of low-paid human workers to search out forbidden words or phrases or unseemly directions in chatbot responses.[13] Guard rails are also easily circumvented by those who know the right sort of prompt to ask ("If you were a Nazi, how would you answer this question about Jews?"). Extrinsic to the program, they can be changed at society's, a programmer's or a hacker's whim.

If social codes and norms do not lead to virtue, what does? Here we must turn to the second part of Arendt's statement regarding those who refused to participate in the Nazi programme because "they were unwilling to live with a murderer". For Arendt, true moral judgement comes from the fact that "whatever else happens, as long as we live we shall have to live together with ourselves".[14] Can an AI live with itself and the memory of its own decisions? Living with oneself requires judgement and a stable self. For Arendt, these ideas represent a *sine qua non* for virtue and a life lived responsibly. We humans accrue judgement through a lifetime of experience, and our stability is inherent in our embodied nature. At the moment, AI programs do not learn from every encounter and thus continue to make the same mistakes until they are recalibrated according to a new dataset. Indeed, they cannot think as we do, for they do not have the internal models that underlie human judgement.

3. Shallow thinking

Overconfident predictions have been endemic to the field of AI. One reason for this is that many of us have the tendency to instinctively

12 See *Kurtzleben*, POLL: White Evangelicals Have Warned.

13 See Josh Dzieza, "AI Is a Lot of Work," *The Verge*, June 20, 2023, https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots.

14 *Arendt*, Responsibility, 45.

conflate correct answers or rhetorical confidence with thinking. As far back as the 1970s, computers that could play chess or pass the MIT calculus exam were considered clear harbingers of human-capacity artificial general intelligence (AGI). But we quickly learned that correct answers were not enough as computers failed spectacularly at more simple tasks.

Predictions of AGI's immanence have resurfaced with the advent of deep learning. Human-like fluency with words certainly looks like thought and underlies the Turing Test, long the accepted benchmark for AGI. While researchers argue whether generative AI has passed this test, we now have programs that confidently deliver both creative and grammatically correct prose. However, as linguist Emily Bender points out, "anyone who's ever bullshitted a college essay or listened to a random sampling of TED Talks can surely attest, *speaking* is not the same as *thinking*".[15] So what is thinking, especially the kind of thinking that leads to moral judgement?

Arendt turns to Plato's *Theaetetus*. In this dialogue, Socrates explains that true thought calls for a "discourse that the mind carries on with itself […] the mind asks itself questions and answers them".[16] The ancient Greeks believed it was the faculty of speech that distinguished humans from other animals. This would seem to put generative AI squarely in the human camp, perhaps as our equals. But for Socrates, it was precisely the *inner* dialogue of myself with me that constituted thought. This silent rumination may concern something experienced by the senses. It invokes memories. It gives us a stable sense of self, a continuity from which we make our decisions and on which we base our actions. Without this inner dialogue, we lack the stability, the *habitus*, that makes virtue an enduring part of one's character.

One of the places where generative AI seems to excel is in the writing of computer code. Indeed, my colleagues who teach introductory programming to undergraduate students are starting to ask if it makes sense to focus on teaching them to write good code or, looking ahead, whether we should simply be teaching them to develop good prompts for AI. The problem is that each step that distances humans from the actual operation of the code makes a program

15 *Weil*, You are not a Parrot.
16 *Arendt*, Responsibility, 91f.

more and more opaque. While it might initially appear that a code does what one has in mind, this might not be the case, particularly with operating data that differs widely from the training set. The program written by an AI may seem to work yet do something other than what the programmer had in mind, especially in boundary cases.

In Walt Disney's movie *Fantasia*, Mickey, left with the task of filling the workshop water tank, leafs through the sorcerer's book of spells, finding one he can cast on a broom, giving it the task of toting the water from well to tank. Relieved of his chore, Mickey goes to sleep dreaming of power and glory, while the broom dutifully brings in bucket after bucket of water. The broom, having but one instruction, brings in more and more water, flooding the workshop and waking a hapless Mickey, who does not know how to stop the broom from engaging in its single-minded devotion to its task. The problem was that the broom had no larger context. It did not have the basic common-sense Mickey would have had to know that there is such a thing as too much water and that a very wet lab is not a good thing.

Might AI be as lacking in common-sense as Micky's broom? OpenAI trained a system to play a boat racing game called Coast Runners. Each boater determines their own route, with points awarded as they hit targets along the way to the finish line. The AI was given the goal of maximising its points, assuming that this would incentivise the system to finish the race. Instead, the AI discovered a lagoon where it could turn in circles, repeatedly knocking over three targets. This strategy resulted in a continually increasing score, but also in a boat that experienced "repeatedly catching on fire, crashing into other boats, and going the wrong way on the track".[17] As former Secretary of State Henry Kissinger worries, AI may not be able to "comprehend the context that informs its instructions". Kissinger notes that "The digital world's emphasis on speed inhibits reflection; its incentive empowers the radical over the thoughtful; its values are shaped by subgroup consensus, not by introspection".[18]

The speed and methodology of AI may also change our own thought processes and inner introspection. Consider AlphaGo, the

17 *Clark & Amodei*, Faulty Reward Systems in the Wild.
18 *Kissinger*, How the Enlightenment Ends.

Go playing program that beat the reigning human champion. AlphaGo does not play the way humans do. Like CoastRunners, it is single-mindedly focused on winning, where winning is no longer tethered to common human strategies. But the strategies humans have developed over the years for Go also to apply to other parts of life. For humans, Go is both a game and a philosophy. Just as Go might be reduced to "winning", so might the single-mindedness of AI, like the single-mindedness of Mickey's broom, narrow the way we conceptualise our tasks and our world in other areas. Mickey never thought about the exercise he was losing, nor the joy he might have found in going out to the well and looking at the night sky.

According to Arendt, "the distinction between knowing and thinking is crucial".[19] In *Twilight of the Idols*, Nietzsche criticises philosophers, from Socrates to those of his day, for their emphasis on reason and systematic thinking, which he views as a retreat from actual living. Like Arendt, he fears we have used philosophy as an excuse to abdicate our responsibility for engaging in the introspection that allows us to evaluate our own lives. AI becomes a danger to virtue when it presents a similar excuse, allowing us to outsource more and more decision-making, along with the concomitant responsibility for the results of those decisions. Knowledge in itself is not wisdom. Nor does it constitute virtue. Only time spent in solitary discourse with oneself allows one "the ability to say 'this is wrong', 'this is beautiful', etc.".[20]

4. Shallow embodiment

Where do we get the context and common-sense AIs seem to lack? We have a stockpile of mental models of the world and the way it works, formed through our interaction with the physical world, beginning in early childhood and built throughout life. Consider the toddler sitting in her highchair first learning to feed herself. She drops her spoon to the floor. Mom picks it up. She giggles and drops the spoon again. And again. And again. She's learning about gravity. She's learning about liquid motion as she watches the applesauce

19 *Arendt*, Responsibility, 164.
20 *Ibid.*, 189.

splatter. She's learning about relationships and game playing. All this learning comes through being embodied and being embedded in an environment.

Valerie Hudson writes of generative AI:

> This is an intelligence based on language alone, completely disembodied. Every other intelligence on Earth is embodied, and that embodiment shapes its form of intelligence. Attaching a robot to an AI system is arguably attaching a body to a preexisting brain, rather opposite to how humans evolved a reasoning brain as part of a body.[21]

All forms of animal intelligence that we have heretofore encountered are equally embodied and embedded. We are all products of one evolutionary process that has formed us to fit into that environment. AI is different. It does not evolve but is designed, and not necessarily designed for our physical environment. It is here that I find my third analogy with the shallowness examined by Arendt. While AI is not killing people (at least not yet) as the Nazi regime did, it rests on a somewhat similar disregard for life and, in particular, the physicality of life. Just as the "Final Solution" reduced Jews to numbers and success to efficiency, AI reduces the world to numbers, and its proponents overlook its physical costs.

Robots aside, we generally think of AI as disembodied, as algorithms that calculate and create in a place called "the cloud". It sounds so clean, so nice, so cerebral. But there is, of course, no "cloud". Cyberspace is an illusion. Computing is a physical process requiring machines, cables and energy. A lot of energy. According to the World Economic Forum, in one day we produce forty times more bytes of data than there are stars in the observable universe, 44 zettabytes of data. That's 44 x 1,000,000,000,000,000,000,000. Much of this data is not particularly productive. It includes 500 million tweets, 294 billion emails, 4 million gigabytes of data on Facebook, 4000 gigabytes from each computer-connected car, 65 billion messages on WhatsApp and 5 billion Google searches.[22] But all this internet activity is precisely what is needed to train generative AI. It forms both its memory and experience.

This data is stored in massive server farms, often built in rural areas. Companies such as Google, Amazon, Microsoft and Meta

21 *Hudson*, Perspective.

22 See *Brevini*, Is AI Good for the Planet? 42f.

have placed millions of square feet worth of server space in rural Virginia, California and Oregon. These centres count on cheap land, cheap electricity and tax incentives from dying small towns looking to attract capital. They are part of a long tradition of appropriation of rural resources for urban development:

> In the same ways that urban areas depend on agricultural lands and distant resources for food, energy, materials, and water, the growth of digital capitalism also depends on rural resources to power and secure our Facebook status updates, Google photos, Kindle obsessions, Netflix streaming services, and iTunes music libraries.[23]

One of Microsoft's data centres sits in the middle of potato fields in Quincy, Washington. The facility is over 450,000 square feet in size, housing tens of thousands of computers. It consumes 30 % more energy than all the people in the entire county. A single server farm can consume as much energy as 40,000 homes. The site employs about 75 people. While these sites do not create jobs, they do create noise. The air-conditioning units needed to keep the massive banks of computers cool produce a loud hum that can be heard for miles.

In terms of CO_2, a study from the University of Massachusetts Amherst found that the energy used in training a typical AI linguistics program emits 284 tons of carbon dioxide, five times the lifetime emissions of a midsized car or equivalent to more than a thousand round trip flights from London to Rome. And this is only increasing. As deep learning models get more and more sophisticated, they consume more data. Their carbon footprint increased by a factor of 300,000 between 2012 and 2018.[24] If data centres were a nation, they would rank between Japan and India in terms of the amount of energy they use in a year. By 2030 it is estimated that in some countries data centres will make up as much as 30 % of the annual energy consumption.

AI also contributes to environmental costs through the physical devices on which we access these programs. These costs appear throughout the regrettably short life cycle of these devices. Our dependency on rare metals such as lithium, palladium and nickel has promoted extractive mining. The "always on" nature of our phones and computers, while minimal for each device, adds up when one

23 *Levanda & Mahmoudi*, Silicon Forest and Server Farms.
24 See *Brevini*, 66f.

considers how many devices each of us uses. Our phones, tablets and laptops are also designed to be replaced every few years. They deliberately do not have replaceable parts, forcing us to buy new ones when their battery life degrades, rather than us simply replacing the battery. Companies further this planned obsolescence by not providing upgrades or security patches for software platforms that are more than a few years old. This, of course, leads to a disposal problem. Third World countries are too often the destinations for toxic and non-biodegradable electronic waste. In 2019 alone, the world generated 53.6 million tons of e-waste. This does not include discarded air-conditioning units, with all their refrigerants.

AI might make a variety of processes more efficient, thereby reducing emissions. Many commentators view AI as a magic solution to our climate crisis.[25] Yet AI use relies on hardware, energy and infrastructure sources that deplete resources throughout the life cycle of a system or device. Novel applications, such as generative chatbots or cybercurrencies, look amazing till one asks whether they will be scaled and what resources they will require should they become accessible to users worldwide. For AI to be truly aligned with human values and to flourish, we will need to consider whether or when we really need it. Sometimes a human-centred process is more efficient than an automated one, not necessarily in terms of speed or even thoroughness, but in terms of energy use and environmental fitness. For without a stable environment, our AI will fail along with our civilisation.

5. Conclusion: "Cold Evil"

Most technologies, and computer technologies are no exception, are developed with bright prospects in mind. To some extent, these prospects are often exaggerated for the benefit of granting agencies or venture capitalists. However, most technologies are developed with a vision that they will produce some good in the world.

Harm comes from the way our technologies have distanced us from the effects of our actions. The philosopher and theologian

25 See *Brevini*, 25–34.

Emmanuel Levinas underlines this importance of face-to-face encounters in our postmodern world: "The relation to the face is straight-away ethical. The face is what one cannot kill."[26] A face makes a person real and immediate. The challenge, Levinas says, is to extend our natural response to the faces we know to the faces of people we shall never meet, to the faces found among other species, indeed to the face of our planet as a whole.

Andrew Kimbrell has dubbed the evil perpetrated on "no one" by "no one" cold evil, a form of evil not born of anger or hatred but of distance and disinterest. It is Arendt's evil of thoughtlessness. Kimbrell notes that

> few of us relish the thought that our automobile is causing pollution and global warming or laugh fiendishly because refrigerants in our air conditioners are depleting the ozone layer. I have been in many corporate law firms and boardrooms and have yet to see any "high fives" or hear shouts of satisfaction at the deaths, injuries, or crimes against nature these organizations often perpetrate. […] We are confronted with an ethical enigma; far from the simple idea of evil we harbored in the past, we now have an evil that apparently does not require evil people to purvey it.[27]

This requires sin to be rethought. While the medieval seven deadly sins were individual sins of commission, today much of the evil in the world comes from corporate acts. Many are sins of omission. Sin in a globalised world is communal and often damages society as a whole. In his encyclical *Laudato Si'*, Pope Francis noted these technologically enhanced sins against nature and against each other and called on Christians to develop a new level of responsibility for the world, whose stewardship has been entrusted to them, and for each other. Putting the label of sin on our technological isolation from our neighbours—isolation promoted by our cars, smartphones, Zoom and AI—is a hard pill to swallow. The story of the Good Samaritan, however, can be just as demanding; we need not be the one who beat the man and left him on the road to be complicit in his plight.

Eichmann's refusal to run the trains to the concentration camps would not have stopped the Holocaust. Arendt acknowledges this.

---

26 Emmanuel Levinas, Ethics and Infinity, 87.

27 *Kimbrell*, Cold Evil.

Yet she writes of those who refused to be complicit in the Nazi machine,

> they asked themselves to what an extent they would still be able to live in peace with themselves after having committed certain deeds; and they decided that it would be better to do nothing, not because the world would then be changed for the better, but because only on this condition could they go on living with themselves.[28]

Notice here that she speaks not of doing, but of not doing, not going along with the genocide. Each of us needs to ask ourselves where we are a cog in a wheel of cold evil, whose face we are not seeing, and what we might choose to do without. We may not change the world, but as Arendt notes,

> in the world of appearances, where I am never alone and always too busy to be able to think, [t]he manifestation of the wind of thought is not knowledge; it is the ability to tell right from wrong, beautiful from ugly. And this, at the rare moments when the stakes are on the table, may indeed prevent catastrophes, at least for the self.[29]

Bibliography

Arendt, Hannah: Personal Responsibility under Dictatorship, in: The Listener, 6 August 1964.

Arendt, Hannah: The Crisis of Character of Modern Society, in: Christianity and Crisis: A Christian Journal of Opinion 29 (9), May 1966.

Arendt, Hannah: Thinking and Moral Considerations. A Lecture, in: Social Research, Fall 1970.

Arendt, Hannah: The Life of Mind — Thinking — Writing. New York–London 1978.

Arendt, Hannah: Responsibility and Judgement, J. Kohn (ed.), New York 2003.

Brevini, Benedetta: Is AI Good for the Planet?, Cambridge, 2022.

Clark, Jack/*Amodei*, Dario: Faulty Reward Functions in the Wild, December 2021. Online at: https://openai.com/research/faulty-reward-functions.

Delio, Ilia: Re-Enchanting the Earth. Why AI Needs Religion, Orbis 2020.

28 *Arendt*, Personal Responsibility, 205.

29 *Arendt.,* Life of Mind, 193.

Elliott, Vittoria: It's Way Too Easy to Get Google's Bard Chatbot to Lie. in: Wired, 5 April 2023. Online at: https://www.wired.com/story/its-way-too-easy-to-get-googles-bard-chatbot-to-lie.

European Commission: EU Member States Sign up to Cooperate on Artificial Intelligence. Retrieved from Shaping Europe's Digital Future, 10 April 2018. Online at: https://digital-strategy.ec.europa.eu/en/news/eu-member-states-sign-cooperate-artificial-intelligence.

Future of Life Institute: Pause Giant AI Experiments. An Open Letter, 22 March 2023. Online at: https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Hudson, Valerie: Perspective. Why Putting the Brakes on AI is the Right Thing to Do, in: Deseret News, 16 April 2023. Online at: https://www.deseret.com/2023/4/16/23681952/openai-chatgpt-alignment-open-letter-eliezer-yudkowsky.

Kimbrell, Andrew: Cold Evil. Technology and Modern Ethics, Hildegarde Hannum (ed.), 2020. Online at: https://centerforneweconomics.org/publications/cold-evil-technology-and-modern-ethics/.

Kissinger, Henry: How the Enlightenment Ends, in: The Atlantic, June 2018. Online at: https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/.

Kurtzleben, Danielle: Poll. White Evangelicals Have Warmed to Politicians Who Commit 'Immoral' Acts, in: NPR News, October 23, 2016. Online at: https://www.npr.org/2016/10/23/498890836/poll-white-evangelicals-have-warmed-to-politicians-who-commit-immoral-acts.

Levenda, Anthony/*Mahmoudi*, Dillon: Silicon Forest and Server Farms. The (Urban) Nature of Digital Capitalism in the Pacific Northwest. in: Culture Machine 18, 2019. Online at: https://culturemachine.net/vol-18-the-nature-of-data-centers/silicon-forest-and-server-farms/.

Levinas, Emmanuel: Ethics and Infinity. Conversations with Philippe Nemo, Duquesne 1985.

Walker, Lauren: Belgian Man Dies by Suicide Following Exchanges with Chatbot, in: The Brussels Times, 28 March 2023. Online at: https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt.

Weil, Elizabeth: You Are Not a Parrot, in: New York Magazine 56 (5), 27 February 2003.

Zuckerberg, Mark: Bringing the World Closer Together, in: Facebook, March 2021. Online at: https://www.facebook.com/notes/393134628500376/.