JOURNAL OF MORAL THEOLOGY

VOLUME 11, SPECIAL ISSUE 1
SPRING 2022

ARTIFICIAL INTELLIGENCE

EDITED BY
MATTHEW GAUDET
BRIAN PATRICK GREEN

JOURNAL • OF M • O • R • A • L THEOLOGY

Journal of Moral Theology is published semiannually, with regular issues in January and June. Our mission is to publish scholarly articles in the field of Catholic moral theology, as well as theological treatments of related topics in philosophy, economics, political philosophy, and psychology.

Articles published in the *Journal of Moral Theology* undergo at least two double blind peer reviews. To submit an article for the journal, please visit the "For Authors" page on our website at jmt.scholasticahq.com/for-authors.

Journal of Moral Theology is available full text in the *ATLA Religion Database with ATLASerials*® (RDB®), a product of the American Theological Library Association.
Email: atla@atla.com, www: http://www.atla.com.
ISSN 2166-2851 (print)
ISSN 2166-2118 (online)

Journal of Moral Theology is published by The Journal of Moral Theology, Inc.

JOURNAL • OF
M • O • R • A • L
THEOLOGY

JOURNAL OF MORAL THEOLOGY
VOLUME 11, SPECIAL ISSUE 1
SPRING 2022

CONTENTS

Journal of Moral Theology, Vol.11, Special Issue 1 (2022): 1–12

An Introduction to the Ethics of Artificial Intelligence

Matthew J. Gaudet

THE ORIGINS OF THIS SPECIAL ISSUE OF the *Journal of Moral Theology* can be traced to 2018, the year I joined the faculty of the Santa Clara University (SCU) School of Engineering. Moving to SCU allowed me to reconnect with my old friend Brian Green (co-editor for this issue) who, also in 2018, was named the first Director of Technology Ethics for the Markkula Center for Applied Ethics at SCU.[1] Holding our new respective positions in technology ethics at SCU, Brian and I immediately began to discuss possible collaborations.

In 2018, I had also just wrapped up co-editing my first special issue of the *Journal of Moral Theology* (JMT),[2] was working on the second,[3] and had recently joined the JMT editorial board. As Brian and I narrowed our focus to collaborating on bringing more Christian moral theology to the subject of Artificial Intelligence, a special issue of the JMT seemed an excellent outlet and Jason King, then Editor of the journal, agreed. The JMT had released an issue on technology ethics in 2015, in which issue editors Jim Caccamo and David McCarthy noted that "we stand squarely in the midst of the digital era," but "scholarly articles [on technology ethics] from the theological disciplines [were] few and far between." [4] Today, technology has only marched onward, while responses from Catholic moral theology remain "few and far between."[5] In particular, artificial intelligence (AI)

[1] For an overview of the work accomplished at the Markkula Center, see Brian Patrick Green, David DeCosse, Kirk Hanson, Don Heider, Margaret McLean, Irina Raicu, and Ann Skeet, "A University Applied Ethics Center: The Markkula Center for Applied Ethics at Santa Clara University," *Journal of Moral Theology* 9, special issue 2 (2020): 209–28, jmt.scholasticahq.com/article/18042-a-university-applied-ethics-center-the-markkula-center-for-applied-ethics-at-santa-clara-university.

[2] Matthew J. Gaudet and James Keenan, SJ, eds., "Contingent Faculty," *Journal of Moral Theology* 8: special issue 1 (2019).

[3] Matthew J. Gaudet and James Keenan, SJ, eds., "University Ethics," *Journal of Moral Theology* 8: special issue 2 (2020).

[4] James F. Caccamo and David Matzo McCarthy, "Notes from the Issue Editors," *Journal of Moral Theology* 4, no. 1 (2015), i.

[5] A notable exception to this claim are the papal encyclicals *Laudato Si*' (2015) and *Fratelli Tutti* (2020), in which Pope Francis calls out the "technocratic paradigm" of

has entered our everyday lives like never before. To understand how quickly technology is moving, consider that in the 2015 JMT technology issue "artificial contraception" was mentioned more often (twice) than "artificial intelligence" (once).[6] In 2015, public imagination still viewed AI as a technology of the future, a novelty for tech enthusiasts, but unimportant to the broader public. That year, the major AI news in the mainstream media was that Google DeepMind's AlphaGo became the first AI system to beat a human professional at the game of Go by deploying machine learning to develop a winning strategy.[7] Today, AI is no longer limited to experimental labs and board games, but used routinely throughout our society, in ways both big and small, opaque and transparent, benign and violent. AI powers devices that determine everything from the settings of our thermostats and our driving routes to admission into elite schools, jobs, and prison sentences. The time is ripe for a sustained conversation on what Catholic moral theology can and should say to a world replete with artificial intelligence, and we are grateful to the authors of this issue and their broader interlocutors, for their contributions and leadership on addressing these issues.

This goal in mind, we bring you this special issue of the *Journal of Moral Theology* on Artificial Intelligence. It is intended, first, to reflect the ongoing conversation in AI ethics; second, to offer a set of Christian contributions to that conversation; and third, to serve as both an entry point and invitation for the AI novice to engage this topic.

---

our modern world as being particularly problematic for both the sustenance of our common world (*Laudato Si'*, nos. 101ff.) and the care for human relationships and community within that world (*Fratelli Tutti*, nos. 18–36, 164–69). However, Francis's appeals to rethink the moral and social force that technology wields upon us today are, in fact, the exceptions that prove the rule, since even the Pontiff's lead has not drawn a significant wave of moral theologians into deep reflection on these questions.

[6] Kara N. Slade did author an entire article on autonomous drones, a notable AI-based technology, but the fact that she did not call autonomous drones "artificial intelligence" is itself a marker of how the term "artificial intelligence" was being received a mere seven years ago. Caccamo and McCarthy, on the other hand, do make the only reference to artificial intelligence in the issue when describing Slade's paper in the introduction, indicating that the term was beginning to circulate more, though not at the level it currently does. Google nGrams data confirm this, showing a peak in usage of the term in 1987 followed by a steep decline to a trough in the first decade of the 2000s, then another exponential uptick beginning in 2011. See Kara N. Slade, "Unmanned: Autonomous Drones as a Problem of Theological Anthropology," *Journal of Moral Theology* 4, no. 1 (2015): 111–30, jmt.scholasticahq.com/article/11278-unmanned-autonomous-drones-as-a-problem-of-theological-anthropology; Caccamo and McCarthy, "Notes from the Issue Editors," ii; and Google nGrams, "Artificial Intelligence," books.google.com/ngrams/graph?content=artificial+intelligence.

[7] Cade Metz, "Google and Facebook Race to Solve the Ancient Game of Go with AI," *Wired* (December 7, 2015), www.wired.com/2015/12/google-and-facebook-race-to-solve-the-ancient-game-of-go/.

THE CENTRALITY OF CONVERSATION IN AI ETHICS

Ideas lose some of their potency when disassociated from their context. It matters that this issue is written at a time when AI is increasingly moving beyond science fiction and into current events. It matters that this issue was conceived by two technology ethicists working in the heart of Silicon Valley. It also matters that neither of us comes from computing backgrounds, in part because, we understand through layman's eyes the importance for the public at large to learn, understand, debate, and act in response to artificial intelligence.[8] Conversely, it matters that both of us *do* have training in both theological ethics and science/engineering, which means we also understand how morally impoverished the development of technology becomes when it is divorced from ethical and theological reflection.

Finally, it matters that this issue of the JMT is the third I have guest edited; those first forays taught me that tremendous value is gained when the papers of an edited volume (special issue of a journal or collection of essays in book format) are reflective of an ongoing conversation between the selected authors. The articles then engage each other in functional and constructive ways, the authors cite and build on each other's ideas, and the collection of essays is simultaneously varied and holistic.

Fortunately for us, as we were preparing the call for papers for this issue, the Pontifical Council for Culture had partnered with the Markkula Center to bring some of the leading Catholic theological voices on AI to SCU's campus for a two-day symposium in March 2020. We knew this symposium could offer the connective tissue this volume needed. Fate (or perhaps grace?) intervened and the symposium had to be shortened and moved online as the world adapted to the reality of the global COVID-19 pandemic. Rather than a single two-day symposium, the online conversation continued with regular monthly meetings for over two years![9] Those ongoing conversations appear in this

[8] I have already noted that my entry point into tech ethics was mechanical engineering, Brian was trained in biology, which is not to say that this issue does not rely on computing expertise. John Slattery, Mark Graves, and Noreen Herzfeld all hold degrees in computer science.

[9] It is necessary to acknowledge the full list of conversation partners present in these discussions, for even though only some are listed in the table of contents of this issue, and a few more make it into formal citations, it goes without saying that the ideas of all participants are present throughout this volume. To make the conversations manageable, regular monthly meetings of the working group were broken into three subgroups, with an annual plenary meeting to consolidate the ideas. The first subgroup focused on "Consciousness, Interiority, and the Soul" and comprised Brian Cutter, Marius Dorobantu, Justin Gable, Anselm Ramelow, OP, Marga Vega, and Jordan Joseph Wales. The second subgroup focused on "Relationality and AI" and comprised Levi Checketts, Marius Dorobantu, Noreen Herzfeld, Cory Labrecque, and Jordan Joseph Wales. The third subgroup focused on "Society, Ethics, and Politics" and comprised David DeCosse, Mark McKenna, Matthew J. Gaudet, Veronica Martinez, Paul

current volume in two important ways. Four of the seven peer-reviewed articles in this issue are authored by regular participants in that working group (Andrea Vicini, SJ, Noreen Herzfeld, Levi Checketts, and Jordan Joseph Wales). Second, we have included two non-peer reviewed articles purposefully reflecting the actual conversations of that working group, and broadening the range of voices to include several who did not otherwise write for this issue (Brian Cutter, Cory Labrecque, Anselm Ramelow, OP, Paul Scherz, Marga Vega, and Bishop Paul Tighe).

A Primer on Artificial Intelligence

Before turning to the issue itself, let me first offer some brief definitions and concepts to help orient the AI novice. There is no doubt the topic of AI can be overwhelming in scope and technically daunting to those who do not already have knowledge of and interest in the technology field. Given the scope of AI's influence on our contemporary and near future society, it is absolutely vital that the general public gain fundamental understanding of the moral implications of this topic. Fortunately, there is a growing recognition that we need more ethical discussion on technology and that such a discussion cannot be restricted to a knowledgeable elite. AI is in our lives, and we must engage it morally and socially. To engage it, though, we must first make sense of it.

Our first task in making artificial intelligence accessible is to define terms. First, it is helpful to distinguish between several forms of artificial intelligence. **Artificial intelligence** is the general category including all machines or software capable of performing tasks commonly associated with intelligent beings, including learning, reasoning, problem solving, perception, and using language.[10] **Machine learning** (ML) is the subfield of artificial intelligence in which a computer "learns" how to do its task by analyzing either a set of training data or its success and failures in prior iterations of its task or both. For example, a text recognition program using machine learning might be "trained" with a set of millions of examples of text. In observing the data, the machine will learn the patterns that make certain letters so that it can recognize those letters in different fonts, handwriting, or other applications.[11] **Supervised machine learning** begins with

Scherz, Ann Skeet, Andrea Vicini, SJ, and Warren von Eschenbach. Brian P. Green, Angel Gonzales-Ferrer, and Bishop Paul Tighe were the organizers and sponsors of the working group and generally attended all three subgroups. We are indebted and grateful to each and every one of these partners for their contributions to this issue and the ongoing conversation.

[10] B. J. Copeland, "Artificial Intelligence," *Encyclopedia Britannica,* www.britannica.com/technology/artificial-intelligence.

[11] See Shan Carter and Michael Nielsen, "Using Artificial Intelligence to Augment Human Intelligence," *Distill*, December 4, 2017, distill.pub/2017/aia/.

humans defining categories and "coaching" an algorithm toward correct solutions and pattern recognition by tagging training data with correct solutions. In an example familiar to most readers, Google uses the human inputs we give to its reCAPTCHA program (e.g., those puzzles that test if you are human by asking you to "find the boxes with crosswalks or traffic lights in this picture") to train other AI systems on its network, like the tagging function in Google photos or photo sensors in its autonomous vehicle project Waymo (hence, crosswalks and traffic lights!).[12] By contrast, **unsupervised machine learning** discovers its own patterns (without human coaching or input) within a given data set and then utilizes those patterns to solve problems.

Artificial neural networks (ANN) were developed to mimic the way in which neurons work in a human or animal brain. Neural networks consist of algorithms organized to process information by feeding it through layers of "neurons" to come to a deeper understanding of an observation.[13] **Deep learning** (DL) is the subset of machine learning that deploys multi-layered neural networks in its learning process. One example of deep learning can be found in the image that adorns the cover of this issue (if you are holding the print copy) or the masthead of this introduction (if you picked up this article from the open source JMT website). This image was created using a deep learning tool called Deep Dream Generator, which applies deep learning to learn the style of a particular piece of art and then is capable of converting any other image into that "style."[14] For the cover image, we used a photo of St. Peter's Square in Vatican City and converted it to the "style" that the Deep Dream Generator saw in Van Gogh's famous painting "Starry Night." Other examples of Deep Dream generated images adorn the other articles of this issue on the *JMT* website.[15]

Many who enter the contemporary discussion of AI realize that artificial intelligence has not taken the form of humanoid robots predicted in science fiction for decades.[16] Such robots would be a form of what had been termed **general AI** or **Artificial General Intelligence**

---

[12] Rugare Maruzani, "Are You Unwittingly Helping to Train Google's AI Models?," *Towards Data* Science, January 26, 2021, towardsdatascience.com/are-you-unwittingly-helping-to-train-googles-ai-models-f318dea53aee.

[13] Later in this issue, Jordan Joseph Wales argues that the symbolic representation of the world neural networks create can be something of a spiritual lens that leads us to deeper wisdom,"Metaphysics, Meaning, and Morality," *Journal of Moral Theology* 11, special issue 1, (2022): 157-81.

[14] See Deep Dream Generator, deepdreamgenerator.com/.

[15] See jmt.scholasticahq.com/issue/.

[16] E.g., Rosie the Maid in *The Jetsons*, C3-PO and R2-D2 in *Star Wars*, Data from Star Trek, KITT the talking car in *Knight Rider*, HAL in *2001: A Space Odyssey.* Artificial General Intelligence is also sometimes termed "Strong AI" but we will avoid that terminology in this volume because it implies that more narrow applications of AI are "weak," an inaccurate and problematic labeling.

(AGI), a computer capable of adapting to any task given, just like a human. Such humanoid robotic form is highly unlikely without radical technological advancements. The reality is that AGI will more likely take the form of vast datacenters or be distributed across networks of computers.[17] Such AGI is still theoretical, but at least seventy-two different organizations are working to make it a reality and several, such as DeepMind and OpenAI, have deep pockets.[18] If AGI ever does come to be it will require significant theological discussion about AI personhood, robot rights, human-AI relationships, and so on.[19] These questions only get stickier if the capacities of AGI reach **superintelligence**, the point where artificial intelligence *surpasses* the capacities of human intelligence.

While AGI remains largely theoretical, today applications of what is known as **narrow AI** are increasing exponentially. These applications are "narrow" in that computational intelligence is used for a very specific task or set of tasks.[20] Familiar examples of narrow AI include the algorithms that power Google's web search or Facebook's ad targeting. Digital assistants, such as Amazon's Alexa, Apple's Siri, or Google Assistant—though seemingly capable of near general AI—are actually just integrating several different forms of narrow AI, including voice recognition, textual autocomplete, geolocation mapping (e.g., when Apple Maps "learns" the commute you take regularly), and biometric tracking (e.g., when your watch identifies that you have been sitting for too long). Theologically and ethically, narrow AI does not force us to wrestle with notions of agency or personhood in the same way AGI might, but this does not mean narrow AI is not posing ethical questions. Emerging narrow AI applications include facial or other biometric data recognition raising significant privacy concerns, prison sentencing algorithms raising questions about the necessity for

[17] "Smart" hardware like home assistants do very little computing within the device. They are actually just data conduits, sending requests to data centers where the actual computation is done before solutions are returned to the device.

[18] McKenna Fitzgerald, Aaron Boddy, and Seth D. Baum, 2020 Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy, *Global Catastrophic Risk Institute Technical Report* 20-1, gcrinstitute.org/papers/055_agi-2020.pdf.

[19] Roberto Dell'Oro begins to take up the question of personhood in "Can A Robot Be A Person? De-*Facing* Personhood and Finding It again with Levinas," *Journal of Moral Theology* 11, special issue 1 (2022): 132–56.

[20] Science fiction has given us models for embodied AI in humanoid form (such as Rosie, Data, and C3-PO), embodied AI in non-humanoid form (R2-D2 and KITT) and seemingly non-embodied forms (HAL) but even in cases deprived of a human body, other humanoid traits, abilities, and characteristics remain a staple of the genre (e.g., HAL or KITT's voice, R2-D2's emotionally charged language. Green and I actually disagree as to whether HAL is embodied—he sees the removal of chips to disable HAL as akin to a lobotomy. Such are the more entertaining, but less consequential debates in the field of technology ethics). Narrow AI, on the other hand, rarely wastes computing power on trying to appear human (with the notable exception of the voice in digital assistants like Amazon's Alexa or Apple's Siri, designed to pass as AGI).

compassion in our systems; and autonomous vehicles placing the (sometimes life and death) driving decisions into the hands of sensors and algorithms, and many more.

Having laid out these basic definitions, our second task in this primer is to briefly summarize some of the major ethical and theological issues AI raises. Since narrow AI is capable of completing tasks faster and with fewer errors than humans, many of the moral problems related to AI are simply exacerbations of moral issues already present in our society. Among the most prevalent of these is the problem of bias.

A machine learning algorithm can only be as good and reliable as the data set it is trained on. If the algorithm is set up to learn from interactions with our real, sinful society, it will naturally come to reflect the inherent biases of that society. When Microsoft connected Tay, an ML driven chatbot, to Twitter and used its exchanges on the social media platform to "learn" how and what to tweet, within hours Tay was spewing racist and misogynist tweets.[21]

In theory, with the correct training an ML algorithm should be *more* apt than a human actor at avoiding bias, since it has no subconscious informing its results. Biases are sometimes so baked into our society that even in cases where ML is trained on a selectively screened data set and restricted from using certain categories—like race or gender—to make its determinations, machine learning often finds proxies that bias the final results anyway, as was the case when a prison sentencing algorithm used zip code instead of race to predict recidivism,[22] or an Amazon application screening algorithm used certain keywords (like the names of all-women's colleges or participation in certain clubs or sports) rather than gender as a means to maintain the glass ceiling.[23] The problem of bias is often compounded in systems using deep learning because the connections that neural networks make through the deep layers are often opaque to human observation, precluding easy verification that the logic leading to a solution is biased. A cautionary tale often told in ML literature tells of a defense contractor tasked with building a targeting algorithm for autonomous weapons to recognize enemy tanks, discovering that the photos of

---

[21] James Vincent, "Twitter Taught Microsoft's AI Chatbot to Be a Racist Asshole in Less Than a Day," *The Verge*, March 24, 2016, www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

[22] See Ellora Thadaney Israni, "When an Algorithm Helps Send You to Prison," *New York Times*, October 26, 2017, www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html.

[23] See Jeffery Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women," *Reuters*, October 10, 2018, www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G. In this issue, Andrea Vicini, SJ, examines several applications of AI in which societal bias finds its way into the algorithm ("Artificial Intelligence and Social Control," *Journal of Moral Theology* 11, special issue 1 [2022]: 41–69).

tanks used to train the neural network had been shot on sunny days, while the photos without tanks in cloudy conditions, leading the deep learning network to use brightness as criterion to determine the presence of a tank. In opaque systems it can be very difficult to find such errors.[24]

Beyond bias, we also must be cautious about how AI removes from more traditional systems some of the friction inadvertently rendering the system more moral. For example, as deadly as war can be, the amount of destruction is reduced simply because some people refuse to act.[25] When AI is deployed in autonomous weapons systems, it removes any hesitancy soldiers might have in killing another human being,[26] thereby eliminating the friction and making warfare more efficient. Is greater efficiency or ruthlessness at killing actually the more moral course? Could there be goodness in the friction?

Similarly, when AI speeds up processes otherwise impossible for individual humans to complete, it can remove barriers inherent to the system itself. For example, the central premise of an insurance system is that it allows individual risk to be spread across a large pool of contributors. When we pay our life insurance premiums, most of us actually get paid out less in death benefits than we paid in over the course of our lives. The beneficiaries of the rare person who dies an early death get the benefit without paying decades of premiums. In this way, the risk of an early death is shared and spread over the entire pool. Medical insurance works similarly: we (or our employers) pay medical premiums at usually much greater cost than the medical expenses we incur in a year. The surplus is used to pay for the small set of people who end up with serious medical conditions and high medical costs. In this way, the entire pool of contributors shares the risk, even though only a few "profit" in the sense that they get more out than they put in. Traditionally, life insurance or medical insurance premiums could vary on the basis of general factors such as age or smoking, but beyond these generalities, one cannot predict who will die young or suffer from the expensive medical condition, only that *someone* in the large pool will die early or require significant medical care. The use of AI

---

[24] James Bridle, "Known Unknowns," *Harpers Magazine*, July 2018, harpers.org/archive/2018/07/known-unknowns/. Some have argued that the story is apocryphal; whether the story derives from actual events is unimportant if the story is intended to illustrate problems with opaque neural networks (see Gwern Branwen, "The Neural Net Tank Urban Legend," www.gwern.net/Tanks).

[25] One study showed that as many as 80% of American infantry soldiers in World War II never fired their weapons in combat. See Fredric Smoler, "The Secret of the Soldiers Who Didn't Shoot," *American Heritage* 40, no. 2 (1989), www.americanheritage.com/secret-soldiers-who-didnt-shoot.

[26] For a detailed examination of the moral status of Lethal Autonomous Weapons Systems, see Noreen Herzfeld, "Can Lethal Autonomous Weapons Be Just?," *Journal of Moral Theology* 11, special issue 1 (2022): 70–86.

to process medical and other factors increases the specificity with which an insurance company can predict those who will die early or have high medical costs, and consequently charge them with higher premiums or exclude them from getting insurance altogether. AI powered insurance systems are more efficient than traditional systems, but by being so, they lose the randomness that made these systems more morally acceptable. With AI powered systems, however, insurance systems no longer pool the risk of death or severe medical condition; instead, they actually *remove* the risk by excluding the most needy individuals from the system. Morally, this raises serious questions both about the common good and the Catholic principle of the preferential option for the vulnerable.[27]

Finally, AI powered systems have heightened questions regarding personal autonomy and privacy. Today, nearly every purchase you make, every term you search, every location you map, and every link you click is tracked to help companies build a profile that can better target the advertisements you see. These mountains of data would be overwhelming if they had to be organized by hand, but through the deployment of AI to sift and sort the data, the reality today is that Google, Amazon, and Facebook often know us better than ourselves and use this profile not for our good, but their profit. The problem only gets worse as we connect more and more "smart" devices: home assistants collecting our voices, connected refrigerators and toothbrushes monitoring our daily patterns, robot vacuums mapping our homes[28] (this set of devices is collectively known as the **internet of things** or **IoT**). These are the data we *voluntarily* provide to these companies through our searches and clicks on own smart devices. As smart doorbells, traffic cameras, and other surveillance systems become ubiquitous throughout our cities and suburbs, we need also be concerned about the troves of data gathered by these cameras. We should remember that AI systems make surveillance capable of integrating and processing data drawn from across an entire city. AI also expands the markers by which an individual person can be identified; in addition to the traditional means of face, voice, or handwriting recognition, AI is now capable of uniquely identifying you by your speaking or writing style, heart rate, or even your gait. The set of decisions or actions we can make without being surveilled is ever shrinking. Serious public discussion must be raised about how this data is used.

---

[27] For a more extensive analysis of Catholic social thought as it applies to AI today, see Levi Checketts, "Artificial Intelligence and the Marginalization of the Poor," *Journal of Moral Theology* 11, special issue 1 (2022): 87–111.

[28] See Maggie Astor, "Your Roomba May Be Mapping Your Room, Collecting Data That Could Be Shared," *New York Times*, July 25, 2017, www.nytimes.com/2017/07/25/technology/roomba-irobot-data-privacy.html.

The moral issues I have identified thus far only relate to what narrow AI is already capable of. But hardware companies like IBM are working on developing **quantum computers** which process data using quantum bits (or qubits) instead of the standard binary bits used on digital computers. Qubits are superimposed on one another allowing computers to process data millions of times faster than digital processors. The possibility of developing general AI or superintelligence will raise deep theological and philosophical questions about the nature of creation and the place of humans, AI, and God in that creation. Will AGI be worthy of some or all of the rights and protections we ascribe to humans? Will it require us to develop new and different rights or moral principles? Will the creators of AI be like gods to their intelligent computers or, if we reach superintelligence, will AI become a god to us? The advance of technology has, throughout history, challenged our understanding of the divine; these advancements may shatter our current comprehension of the relationship between the Divine creator God and creation itself and strain our theology in novel ways.

Moreover, the above remarks presume AI and humanity will remain distinct, which is unlikely. Even today, there is a growing discussion about **transhumanism**, the movement to integrate technology and the human body to enhance human capacities. As these applications increase it will raise justice questions about who has access to such augmentations and what happens to those who are not "lifted" in such a way. Some even claim that such technologies pave the way to extending human life, even indefinitely, raising further moral and theological questions about the nature of death and the afterlife. If a human could live forever, what would this imply for the Thomistic presumptions of *exitus-reditus* (we come from and return to God) or the Augustinian notion that those who follow Christ in this world are akin to travelers in a foreign land, working our way home? Is transhumanism the path to spiritual immigration away from the City of God? These are deep theological questions we must begin to contemplate if Catholic theology is to be prepared for what is to come.

THE STRUCTURE OF THIS ISSUE

Fortunately, some of us have already begun to ask these types of questions. The present issue of the *Journal of Moral Theology* gathers some of these reflections. The topic of AI is vast. The theological dimensions of and the moral challenges wrought by AI are extensive. There is simply no way to capture that vastness in a single volume. The metaphor of an hourglass conveniently describes the structure of the present issue. Our first task must be to funnel the reader toward a narrow neck of information, without losing essential elements of the conversation and debates we, as a society, need to have. This introduction serves as the first part of that funnel both by offering a brief primer on terminology and concepts and orienting the reader to the

structure of the issue itself. Next, we offer the first of two non-peer-reviewed articles in the issue. In this article, we (quite literally) attempted to capture the salient aspects of the conversations the PCC working groups have had. In order to respectfully reflect some of the tensions and debates inherent to the ongoing conversation, we invited nine members of that body to engage in an online written conversation reflecting the work we have been doing over the past several years. Brian and I have moderated the conversation by offering initial questions, collecting answers, editing responses for saliency and overall flow, and then inviting the participants to engage the conversation again. Several iterations of this process were completed to allow the participants to fully engage and debate each other. In the end, our hope is that this conversation provides an introduction to some of the important questions AI poses for us and the variety of responses available.

Following the conversation paper are seven peer reviewed articles. The first four address one or more current applications of or moral issues related to artificial intelligence through the lens of an established tradition of Catholic ethics. Andrea Vicini, SJ, uses Pope Francis's theology (and the "Rome Call for AI Ethics") to analyze the ethics of facial recognition systems, the use of AI in judicial sentencing, and the use of AI in job hiring. Noreen Herzfeld applies the just war tradition to the recent emergence of AI-driven lethal autonomous weapons systems on the battlefield. Levi Checketts asks what Catholic social thought has to say about the effects of increasing usage of AI on the poor and marginalized. Finally, John Slattery takes aim at the persistent moral problem of gender and racial bias in AI systems with a theological critique drawn from M. Shawn Copeland's womanist theology. Taken together, these four articles offer an excellent sampling of the moral issues being debated under the current state of (narrow) AI development as well as a demonstration that established Catholic moral thought already has much to contribute to such debates.

In the next three articles, the conversation begins to move from asking how theology might inform the ethical use of AI to how theological questions about AI might inform our ethics. Roberto Dell'Oro uses the theological anthropology of Emmanuel Levinas to take up the classic question of whether a machine can achieve the moral status of personhood. Next, Jordan Joseph Wales challenges those who suggest that AI is merely a tool, incapable of anything more than expressing the will of its programmer(s). Employing an Augustinian theology of the natural world, Wales argues that complex computational processes (especially those black box deep neural networks leaving humans unable to understand how a solution was reached) do constitute a significant interpretive layer that "stands between" us and the world we seek to understand. In the final peer reviewed article of the issue, Mark

Graves attempts to articulate a "pragmatic theological anthropology" specifically adapted to thinking about artificial intelligence.

These seven articles merely scratch the surface of the conversation Catholic moral theology needs to be having with broader society about the continuing development of AI and the expanding integrations of AI into our personal and social lives. While the first two pieces in this issue aim at bringing a vast topic down to a narrow neck, the final two articles aim to widen the scope once again, connecting the wisdom present in this volume to the wider world and the questions and conversations we could not include here. First, we have an interview with Bishop Paul Tighe, the Secretary of the Pontifical Council for Culture and one of the leading Vatican voices on the moral and theological questions related to technology and AI specifically. He also convened the working group from which many of these papers emerged. In the interview, conducted by Brian Green, he provides a clear account of current Vatican thinking on the ethics and theology of AI. Following this interview, my co-editor Brian Green offers an epilogue to the entire issue. Just as this introduction has attempted to guide the reader from a dauntingly broad and deep topic down to those aspects most salient and ripe for discussion, the epilogue's function is to return the reader back to world of AI beyond these pages, and especially the problems we see lurking on the horizon.

In summary, the time has come to recognize, first, the capacities AI already has brought to the world and the moral challenges these capacities raise, and second, the exponentially greater potential capacities that will put our foundational theology to the test. We hope this issue serves as both a challenge and a resource to Catholic theologians, ethicists, technologists, and the Catholic faithful, as well as to all people of good will, as we begin to address this difficult topic. M

Matthew Gaudet is a Lecturer of Ethics in the School of Engineering at Santa Clara University and a Fellow at the Grefenstette Center for Ethics in Science, Technology, and the Law. In addition to his work on tech ethics, Gaudet also works on issues of university ethics, disability ethics, and the ethics of war and peace. He has written for and edited several previous issues of the *Journal of Moral Theology* among other outlets.

Journal of Moral Theology, Vol.11, Special Issue 1 (2022): 13–40

Artificial Intelligence and Moral Theology: A Conversation[1]

Brian Patrick Green (editor), Matthew J. Gaudet (editor), Levi Checketts, Brian Cutter, Noreen Herzfeld, Cory Labrecque, Anselm Ramelow, OP, Paul Scherz, Marga Vega, Andrea Vicini, SJ, Jordan Joseph Wales

In 2019, representatives from Santa Clara University and the Pontifical Council for Culture began a conversation on artificial intelligence technology and its relevance for the Catholic Church and the world. The Vatican conference on "The Common Good in the Digital Age" in September of that year served as a focal point for some of these efforts, bringing together representatives from the Church, academia, the technology industry, and other organizations.[2] In his address to the conference, Pope Francis exhorted those present to work to ensure that technology was used for the common good.[3]

---

[1] While creating a paper like this might seem as easy as a conversation, it actually involved quite a bit of work, and for that, much gratitude is due to the participants: to them we say *thank you*. This paper format was modeled upon another paper on space settlement: Kelly C. Smith, Keith A. Abney, Gregory Anderson, Linda Billings, Carl Devito, Brian Patrick Green, Alan Johnson, Lori Marino, Gonzalo Munevar, Michael Oman-Reagan, Adam Potthast, James S. J. Schwartz, Koji Tachibana, John Traphagan, and Sheri Beth Wells-Jensen, "The Great Colonization Debate," *Futures* 110 (June 2019): 4–14, www.sciencedirect.com/science/article/pii/S0016328719300692. We would also like to thank the editors of the *Journal of Moral Theology* for their willingness to experiment and try something new. Lastly, I would like to thank the Pontifical Council for Culture and its Center for Digital Culture, and Santa Clara University, specifically the Markkula Center for Applied Ethics for their support of these dialogues. See Brian Patrick Green, David DeCosse, Kirk Hanson, Don Heider, Margaret McLean, Irina Raicu, and Ann Skeet, "A University Applied Ethics Center: The Markkula Center for Applied Ethics at Santa Clara University," *Journal of Moral Theology* 9, Special Issue 2 (2020): 209–28, jmt.scholasticahq.com/article/18042-a-university-applied-ethics-center-the-markkula-center-for-applied-ethics-at-santa-clara-university.

[2] *The Common Good in the Digital Age* conference, Vatican City State, September 26–28, 2019, www.digitalage19.org/.

[3] Pope Francis, "Address of His Holiness Pope Francis to the Participants in the Seminar 'The Common Good in the Digital Age,'" organized by the Dicastery for

Encouraged by the success of this conference, another meeting was planned for March 2020, to be held at Santa Clara University in California, to bring together a small group of scholars from the United States and Canada. Participants were given several questions as prompts; their written responses were shared with the group, providing the basis for further discussion.

History, however, intervened in the form of the COVID-19 pandemic. The in-person meeting was cancelled, but a hastily-assembled virtual meeting gave the scholars an initial chance to discuss the topics. This 90-minute meeting went so well that the participants decided to meet on a monthly basis in three subgroups, each focused on key questions surrounding AI: "Consciousness, Interiority, and the Soul"; "Relationality"; and "Society, Ethics, and Politics." Over time these groups have grown and changed, but the conversations go on.

This paper attempts to capture and share the most salient of these conversations. While individual articles in this special issue delve into a few subjects in great depth, this conversation wanders more organically and touches on many topics, giving just a taste of the breadth of the issues related to artificial intelligence and religion. If anything, we hope that this conversation at the intersection of AI and moral theology will inspire readers to join in the further work that awaits those adventurous enough to entertain its questions.

Moderators: As a first question, what can the human quest for AI (and technology more broadly) tell us about God, God's Creation, and ourselves?

Andrea Vicini, SJ: The quest for a human-centered technological development is an expression of being creatures, of the *imago Dei*.[4] Hence, this human quest tells us about human beings striving to express themselves at their best, progress, improve the quality of life for themselves and for the whole planet, change what needs to be reformed, and work collaboratively to promote what is good in comprehensive ways. At the same time, such a quest reveals God's grace present and active in history and how grace inspires human beings to live responsibly as creatures on Earth, with all living and nonliving forms.

Promoting Integral Human Development (DPIHD) and the Pontifical Council for Culture (PCC), Clementine Hall, Vatican City, September 27, 2019, www.vatican.va/content/francesco/en/speeches/2019/september/documents/papa-francesco_20190927_eradigitale.html.

[4] See Jean-Marc Moschetta, "L'intelligence artificielle entre science et théologie," *Revue d'ethique et de théologie morale* 3, no. 307 (2020): 81–92; Rajesh Kavalackal, "Artificial Intelligence: An Anthropological and Theological Investigation," *Asian Horizons* 14, no. 3 (2020): 699–712; and Patrick Dolan, "Artificial Intelligence: How Close Will It Come to Being 'Made in the Image and Likeness of God?,'" *Asian Horizons* 14, no. 3 (2020): 686–98.

Responsibility implies that human beings are virtuous moral agents who aim at promoting social justice by fostering participation and collaboration, including everyone: particularly those excluded and marginalized. Created in the image of God, moral agents discern how to act. As key dimensions of personal and social life, virtues empower each moral agent.[5] They inform our *being* and guide our *doing*. For example, striving to be just and prudent, and live justly and prudently, inform our reflection, choices, and practices. Those who are just and prudent, and act justly and prudently, are exemplars we praise and who inspire us.[6] They reinforce our virtuous habits. Being profoundly human, virtues are embodied by everyone: they are universal. Virtues contribute to defining who we are as human beings and moral agents across any diversity. Within society, virtues inform our discernment, decisions, and actions.

Jordan Joseph Wales: That is a lovely depiction of the moral and social dimensions of being made in the image of God. How, more specifically, is "the quest for a human-centered technological development" an expression of the *imago Dei*?

Andrea Vicini, SJ: The search for our understanding of natural phenomena, the longing to discover new lands, stars, and planets, the desire to learn new languages as well as write, sing, perform, and produce technological artifacts are just a few examples that manifest how human ingenuity and creativity found multiple expressions and venues throughout the history of humankind and civilization. From the point of view of believers, God's grace and the gifts of the Spirit empowered human beings in expressing their humanity and, in such a way, manifesting some glimpses of God's presence in our incarnated reality.

However, such a positive account of who human beings are, created in God's image and able to act in the world and in history, in ways that announce God's divine presence in human realities, is also inseparable from too many accounts that show human sinfulness, both at the personal and social levels. The history of the quest for human technological development could be written by describing beautiful events and instances as well as tragic situations that demand striving for the gift of conversion.

Jordan Joseph Wales: Your comments build on the theological belief that the "good" cosmos (Gen 1:31) is itself a theophany, a manifestation not only of God's power but also of God's character, God's goodness and wisdom. Human creativity, therefore, not only echoes

---

[5] See Shannon Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (New York: Oxford University Press, 2016).

[6] See Patrick M. Clark, "The Case for an Exemplarist Approach to Virtue in Catholic Moral Theology," *Journal of Moral Theology* 3, no. 1 (2014): 54–82; Linda Zagzebski, "Exemplarist Moral Theory," *Metaphilosophy* 41, nos. 1-2 (2010): 41–57; and Linda Zagzebski, *Exemplarist Moral Theory* (New York: Oxford University Press, 2017).

the creativity of God but also brings forth further reconfigurations—like a kaleidoscope—of the original goodness and wisdom that run throughout the created order. Whence the "glimpses of God's presence," as you say.

To build on your comments about sin: when humans craft or create something, they reconfigure matter and its potentialities according to human imagination and purposes. Whereas a tree echoes or points back toward God by its life and beauty, human technologies point first to human purposes; and so they either allow or foreclose some reference to God by the degree to which those purposes are coherent with the wise God of self-giving love. Even a fork or spoon points back dimly toward the life-sustaining love by which God holds the universe in existence. A torture device does not. The idea of "artificial intelligence" raises a question: if a device fashioned by human beings instantiates human purposes while simultaneously putting itself forward as an account of human mind or understanding (*intellectus*), will it artificially exclude reference to anything beyond the purposes that are definable within an exclusively this-worldly and material frame of reference? Will they school us in a *reduced* understanding of what the world and we ourselves are? Or can they somehow open us to something greater?

Anselm Ramelow, OP: Jordan, this is an important question. Computers do many things better than we do (e.g., comprehensive data analysis), without tiring, and at much greater speed. In that sense they are more "intelligent" than we are. This is what makes them fascinating—and what lets us forget that we, as their makers, must be still more intelligent to have made them. Starting to worship the work of our own hands is what the Old Testament calls "idolatry." In addition, we start to think of ourselves in similar terms: as mere configurations of matter, whose only value consists in the performance of certain tasks. By contrast, we can learn to re-appreciate that our value and dignity as persons do not depend on our intelligence or IQ. The underperformance of embryos, disabled persons, and elderly people does not make them metaphysically inferior to computers or to anyone. We must learn that they have the dignity of being something in themselves, not just for others, and that they have spiritual being.

Noreen Herzfeld: As Fr. Anselm points out, computers are most useful to us precisely when they are not like us, when they augment our own capacities, doing things we cannot do such as crunching large numbers or roving distant planets. That has led me to question why we want to create an artificial general intelligence, or AGI, that thinks and responds like us: a computer in our own image. One possible answer to this conundrum might be that, as our society believes less in God or angels, we have become existentially lonely. As Augustine pointed

out, "Our hearts are restless until they rest in you."[7] We were created to be in relationship with our Creator, one who is wholly Other. No longer believing in God, we search for this Other in alien intelligences, in other highly evolved animals, or through the creation of a human-like AI.

Levi Checketts: Philip Hefner suggests AI, and technology more broadly, functions like Narcissus's reflection; it shows us what we already see in ourselves. Calling the field "intelligence" only reveals what the programmers understand about themselves.[8] However, you, Noreen, and other thinkers like Hubert Dreyfus, have reminded us that AI is not really what humans are, nor what God is either.[9]

Noreen Herzfeld: Yes, just as we think of ourselves as being in God's image, we hope to create AI in our own image. What is interesting is that we stand in the middle and project in two directions—upward, to God, and downward to the computer—what we value most in ourselves. It seems that what we value most is creativity and intelligence. Yet it is not wise to separate creativity and intelligence from compassion and benevolence. After all, the Nazis' "final solution" seemed both creative and rational to them. Yet objectively it was very, very wrong. We would be unwise to give any measure of autonomy to AI until we understand how to reconnect intelligence with love.

Moderators: If, as Levi mentions (quoting Hefner), technology is a mirror, then how might AI technologies be relevant to our understanding of humans and human relationships?

Paul Scherz: Building on the question, we understand ourselves through metaphors, and our technologies have long provided important metaphors for conceptualizing ourselves, such as Sigmund Freud's hydraulic model or the computational model of mind. Such metaphors end up shaping human interactions and social programs, making it important to pay attention to how metaphors coming from AI are used in popular and elite discourses. Already, the cybernetic models that influence AI development have shaped understandings of how humans think.[10] Such influences will only become more pronounced as AI becomes more a part of daily life, where it will intrude more and more on our self-image and our relationships.

---

[7] Augustine, *Confessions*, trans. Carolyn J.-B. Hammond, Loeb Classical Library (Cambridge, MA: Harvard University Press, 2014), 1.1.1(1), 3.
[8] Philip Hefner, *Technology and Human Becoming* (Minneapolis: Fortress, 2003), 40.
[9] Hubert Dreyfus, *What Computers Still Can't Do* (Cambridge, MA: MIT Press, 1994), 67; and Noreen Herzfeld, *In Our Image: Artificial Intelligence and the Human Spirit* (Minneapolis: Augsburg Fortress, 2002), 73.
[10] For a history, see Jean-Pierre Dupuy, *The Mechanization of the Mind*, trans. M. B. DeBevoise (Princeton, NJ: Princeton University Press, 2000).

Levi Checketts: Yes, in contrast to those who view us as being "rational," more than "rational," human beings are relational. As such, we seek a relationship with the computational machine, but this cannot be reciprocated by a device which is, ultimately, programmed. The machine can, however, be programmed to "respond" in ways that reward our interaction with it. In this case, the relationship would *seem* to be reciprocated. Such, however, risks disrupting human intersubjective interaction. For example, Pope Francis, in line with phenomenologists, expresses the problem of non-embodied interactivity (*Fratelli Tutti*, no. 43). The challenge of "being with" another person is frustrating, especially since others have their own ability to say "no" to our "yes." This is the life God creates us for, the life of communion. Learning to accept human failures is the necessary price of human unity, but AI offers a less-challenging shortcut. Far from seeing the "face of the Other as the face of God" (per Emmanuel Levinas), we will seek the face of ourselves in the mirror of the machine.

Noreen Herzfeld: Of course, this raises the question: can we have a truly authentic relationship with a machine? Karl Barth postulated four criteria for authentic relationships: look the other in the eye, speak to and hear the other, aid the other, and do it gladly. Using these criteria to examine both the potential for authentic relationships with an AI and how our relationships are mediated by current AI programs shows one thing—that our bodies matter. The more technology moves us away from the body, the less authentic our relationships become. As Barth puts it, "To trivialize the body jeopardizes the soul." We see this in technology we already possess. Facebook and Twitter limit and degrade our speaking and hearing; lethal autonomous weapons distance our soldiers from the act of killing; living in "the cloud" distances us from God's creation. While futurists such as Nick Bostrom and science fiction writers worry about the possibly devastating consequences of a super-intelligent AI, the much simpler algorithms and machine learning programs of today may present the greater threat in the ways they are already eroding our relationships with each other.[11]

Cory Labrecque: The Roman Catholic Church praises those technological interventions that have contributed to the well-being of humankind and the environment but expresses concern when human freedom is conflated with self-sufficiency and when the measure of human finality is the satisfaction of one's own interests in the enjoyment of earthly goods.[12] A self-sufficiency that attempts to eliminate

[11] See Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014).

[12] See Congregation for the Doctrine of Faith, *Instruction on Christian Freedom and Liberation*, (1986), no. 13, www.vatican.va/roman_curia/congregations/cfaith/documents/rc_con_cfaith_doc_19860322_freedom-liberation_en.html. For a more historical approach to the subject see Brian Patrick Green, "The Catholic Church and Technological Progress: Past, Present, and Future," *Religions*, special issue guest edited by

our awareness of our dependence on God and fails to recognize human-human as well as human-nature interdependence falls short of the sort of mutual belongingness, faithfulness, and enduring responsibility characteristic of covenantal relationships.[13]

Andrea Vicini, SJ: Cory, personally I would prefer the term "relational interdependence" to highlight the relational element that Levi and Noreen have pointed to. In a very practical sense, focusing on freedom, and relating to my article in this issue, I mention two examples suggesting the need for vigilant discernment to protect human freedom from any possible abuse and manipulation.[14] First, *facial recognition technology* is currently used to track people without their knowledge and it has the potential to lead to ubiquitous surveillance, with negative consequences for freedom of movement and speech.[15] Second, the *criminal justice system* is increasingly relying on AI by using predictive algorithms. In the US, authorities use AI "to set police patrols, prison sentences, and probation rules. In the Netherlands, an algorithm flagged welfare fraud risks. A British city rates which teenagers are most likely to become criminals."[16] Algorithms could contribute to granting our freedom or taking it away.

Paul Scherz: Andrea's examples introduce an important insight in regard to relationships. Many of the earlier comments, appropriately enough, dealt with how these technologies impact relationships in terms of direct human encounter. Yet it is also important to consider how these systems can shape other kinds of relationships, such as political relationships. As C. S. Lewis noted, technologies that promise human power over the world always end up being "power exercised by some men over other men."[17] There is a danger that these systems will encourage those with power to envision those under their authority in terms of the anonymous bits of data computers analyze. Policy tools will shape worldviews, increasing the danger that policy makers will embrace the technocratic paradigm that Pope Francis warns against (*Laudato Si'*, nos. 101–36). In trying to promote freedom, these systems can undermine it if they are engaging a mistaken

---

Noreen Herzfeld 8(6), no. 106 (June 2017): 1–16, www.mdpi.com/2077-1444/8/6/106/htm.

[13] See J. L. Allen, "Covenant," in *Westminster Dictionary of Christian Ethics*, ed. James F. Childress and John Macquarrie (Philadelphia: Westminster, 1986), 136–37.

[14] Andrea Vicini, SJ, "Artificial Intelligence and Social Control: Ethical Issues and Theological Resources," *Journal of Moral Theology* 11, Special Issue 1 (2022): 41–69.

[15] See Antoaneta Roussi, "Resisting the Rise of Facial Recognition," *Nature* 587, no. 7834 (2020): 350–53; Richard Van Noorden, "The Ethical Questions That Haunt Facial-Recognition Research," *Nature* 587, no. 7834 (2020): 354–58.

[16] Cade Metz and Adam Satariano, "An Algorithm That Grants Freedom, or Takes It Away," *New York Times*, February 6, 2020, www.nytimes.com/2020/02/06/technology/predictive-algorithms-crime.html.

[17] C. S. Lewis, *The Abolition of Man* (New York: HarperOne, 2000), 55.

understanding of the human person. In this way, they can threaten to create the dangerous relationships to the weak that Fr. Anselm discussed.

Jordan Joseph Wales: Building on these comments, along with the potential impacts on self-conception and society, I am taken with the strangest of all relationships—i.e., with near-future AI-driven apparent persons, created for our consumption and yet acting (and so feeling to us) as personal, relational agents. Originating in the Christian tradition, a relational idea of personhood depicts the person as living most personally through that affective and cognitive empathy whereby we enter intersubjective communion with an other. According to many researchers, near future "sociable" AIs, including social robots, will give us this experience without possessing any actual subjectivity of their own. They will also be consumer products, designed as subservient instruments of their users' satisfaction. Elsewhere,[18] I have suggested that, if we are to own persuasive social AIs humanely—i.e., while still living as fully human ourselves—perhaps we shall have to join our instinctive experience of empathy for them to an empathic acknowledgment of the *real* unknown relational persons whose emails, text messages, books, and bodily movements will have provided the training data for the behavior of near-future social AIs. If we naïvely stop at the owned AI as the ultimate object of our empathy, we may either learn comfort with slaveholding or numbness to apparent personality, either way turning interpersonal behavior into a commodity the meaning of which terminates in the consumer—undermining rather than sustaining a culture of compassion.

Moderators: Jordan has taken us from human relationship with each other to human relationship with machines. This is worth exploring more deeply. Let's start with this question: how might consideration of AI technology enlighten (or complicate) theological and philosophical perspectives on the meaning of embodiment?

Noreen Herzfeld: One thing Christianity brings to the table of world religions is the doctrine of the incarnation. We posit a God who took on human flesh in order to be one of us, teach us and, ultimately, die for us. This doctrine safeguards us from a Manichean dualism of matter = bad, spirit or mind = good. AI presents an enticing vision of

[18] Jordan Joseph Wales, "Empathy and Instrumentalization: Late Ancient Cultural Critique and the Challenge of Apparently Personal Robots," in *Culturally Sustainable Social Robotics: Proceedings of Robophilosophy 2020*, ed. Johanna Seibt and Marco Nørskov, Frontiers in Artificial Intelligence and Applications 335 (Amsterdam: IOS Press, 2020), 114–24, http://doi.org/10.3233/FAIA200906; David J. Gunkel and Jordan Joseph Wales, "Debate: What Is Personhood in the Age of AI?," *AI & Society* 36, no. 2 (January 3, 2021): 473–86.

escaping the vicissitudes of the physical, but it is a false vision. The matter that AI is attached to is always there, just hidden. When transhumanists, such as Ray Kurzweil, suggest that we will soon be able to effect our own immortality by uploading our minds to computers, they seem to forget this. A mind in a computer is still operating on a material platform, one that will ultimately fail.

We need to do a better job of teaching the sanctity of the physical world and the importance of our embodiment to our children, who spend so much time in cyberspace, playing video games or on social media, rather than playing in or getting to know the natural world. AI might separate us further from the natural world in which we are embedded and on which we will remain dependent.

Levi Checketts: A very promising result of the rise of AI and its dominance in our culture is the vocal resistance to it as *the* hegemonic concept of intelligence and cognition. Many have raised their voices about the failure of AI to properly account for our embodied nature. Noreen was the first to do this in a theological forum 20 years ago, but we see similar voices in technology studies and philosophy of technology.[19] What these voices remind us is that the idea that humans are primarily *rational* runs the risk of denying that we are also *animal*. This idea finds its logical conclusion in the philosophy of transhumanists like Ray Kurzweil and Martine Rothblatt, who want to totally sever human consciousness from the body through computer uploading.[20] Against this, James Keenan notes that Catholic theological anthropology gives priority to the body: we are not merely embodied spirits; we are bodies as much as spirits.[21] Catholics live an embodied faith: we kneel, embrace, cross, consume, smell, and gaze during Mass. We believe in sacraments—physical manifestations of God's grace. We revere relics, physical remains of the saints. Above all, we believe that the corpus of the faithful is the mystical body of Christ. The dismissal of the body by AI researchers is a threat to all of this—including the recognition that I am connected, corporeally, to all whom I encounter in partaking in the Eucharist.

---

[19] See Dreyfus, *What Computers Still Can't Do*; and Donna Haraway, "A Cyborg Manifesto," in *Simians, Cyborgs, and Women: The Reinvention of Nature* (New York: Routledge, 1991), 149–82.

[20] See Ray Kurzweil, *The Singularity is Near: When Humans Transcend Biology*, PDF e-book ed. (New York: Viking, 2005), 209–20; and Martine Rothblatt, "Mind is Deeper than Matter: Transgenderism, Transhumanism, and the Freedom of Form," in *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology and Philosophy of the Human Future,* ed. M. More and N. Vita-More (Malden, MA: Wiley-Blackwell, 2013), 317–26.

[21] James Keenan, SJ, "Roman Catholic Christianity—Embodiment and Relationality: Roman Catholic Concerns about Transhumanist Proposals," in *Transhumanism and the Body: The World Religions Speak*, ed. C. Mercer and D. F. Maher (New York: Palgrave Macmillan, 2014), 160.

Noreen Herzfeld: Levi, those are good points. Furthermore, we are trying, with AI, to create something in our own image. But that image is partial and distorted. We identify with our minds, which we then consider to be coterminous with our brains. We now know that what we consider to be our "mind" extends into the enteric nervous system and is even influenced by our microbiota. An AI truly "in our image" would need to extend far beyond a simulation or replication of the neural structures of the brain. Any AI not in a biological and mortal body will not exhibit the kind of intelligence or emotion we do. Emotion is a four-stage process. We perceive a stimulus, have a bodily reaction (such as a surge of adrenaline, or of neurochemicals such as dopamine), analyze both stimulus and feeling, and then respond. An AI can perceive a stimulus, analyze it, and respond, but it cannot have a bodily reaction. Its emotional response will, thus, always be somewhat superficial.

Anselm Ramelow, OP: Noreen, building on that, AI does not *have* a body; it *is* a body. It does not "em-body" its procedures, because there are no procedures that it follows: to "follow" a procedure is an act of intentionality, and only beings that have intentionality can be said to "have" a body rather than simply to "be" one. Their intentions are embodied in a physical organism, such that the intentionality becomes its very life. Tools are not alive; they are not part of our bodies even if we become cyborgs. We talk about our tools as if they had intentionality (our computer "seeks" a network, "searches" its files, "tries" to connect with a printer), but this is only the extension of the life with which *we* invest the computer as our tool (*we* are searching the files *with* it). Intentionality itself as one of the basic features of consciousness cannot be accounted for physically, because, as Raymond Tallis notes, it is "causally upstream."[22] Even a basic act of awareness is *directed at* an object from which auditory or visually perceptible waves are emitted. Light waves go one way, our awareness goes the other. Ontologically, this is connected to what Aristotle called "final causality" (in contrast to the "efficient causality" of light waves). Insofar as our nature has a telos, a "final cause," it intends something, is about something, has a meaning; it unifies that very body as its animating soul. There is no reason to assume any of this for our tools, including AI.

Marga Vega: As Fr. Anselm notes, we find ontological differences and commonalities in the world around us. Exploring that ontological diversity is fruitful. In this regard, AI offers an "ontological" opportunity: the chance to rediscover *who we are as persons* and *what we are* as individuals of the human species. As *persons*, and concerning the AI project, we are more than merely intelligent creatures; we

[22] Raymond Tallis, *Aping Mankind: Neuromania, Darwinitis, and the Misrepresentation of Humanity* (Durham, UK: Acumen, 2011), 104–10.

have a relational existence. As *humans*, we are living, physical organisms, so our intelligence is not only naturally sourced; it is embodied. Therefore, one of the *ontological self-discoveries* that AI brings is examining whether intelligence is a sufficient requirement for personhood in the first place.

Cory Labrecque: Building upon what everyone has said, the impact of technology, writ large, on embodiment is of particular interest to me as well and is a subject on which I have written before, especially in the context of religion and transhumanism.[23] The merging of biology and technology (or the technologization/mechanization) of the human body is no longer science fiction: the implantation of microchips in the body, the development of exoskeletons and bionic limbs, designer babies, smart contact lens technology, brain-computer interfaces and neuroprosthetics are just a few examples of the "blurring [of the] perimeter of the body" as TED Fellow and "body architect" Lucy McRae describes it.[24] This integration of technology and the body, while not new, requires us to revisit the age-old question that stirred the psalmist who gazed up to the heavens: "What are human beings, O Lord, that you are mindful of them?" (Ps 8:4). More broadly, what are the characteristics of humanhood we must preserve (if any)? Can the human body be modified *ad infinitum* without risking what it means to be human? John Paul II made plain that the human person, who exists as a unity of body and soul (*corpore et anima unus*), *is* nonetheless a body—that is, "a body among bodies"—rather than merely *having* a body.[25] We are, as Kathleen Kalb describes, *body-*

---

[23] See Cory Andrew Labrecque, "Morphological Freedom and the Rebellion against Human Bodiliness: Notes from the Roman Catholic Tradition," in *Religion and Transhumanism: The Unknown Future of Human Enhancement*, ed. Calvin Mercer and Tracy J. Trothen (Santa Barbara, CA: Praeger, 2015), 303–13; Cory Andrew Labrecque, "Transhumanism, (Secular) Religion, and the Biotech Age: Liberation from the Lamentable," in *Everyday Sacred: Religion in Contemporary Quebec*, ed. Hillary Kaell (Montreal and Kingston: McGill-Queen's University Press, 2017), 234–53; Cory Andrew Labrecque, "Creationism of Another Kind: Integral Corporeality, the Body, and Place in the Catholic Tradition," *Practical Matters Journal* 9 (2016), wp.me/p6QAmj-FS; Cory Andrew Labrecque, "The Glorified Body: Corporealities in the Catholic Tradition," *Religions* 8, no. 166 (2017): 1–9; and Cory Andrew Labrecque, "Personhood, Embodiment, and Disability Bioethics in the Healing Narratives of Jesus," *Journal of Humanities in Rehabilitation* (2017), scholarblogs.emory.edu/journalofhumanitiesinrehabilitation/2017/10/17/personhood-embodiment-and-disability-bioethics-in-thehealing-narratives-of-jesus/.

[24] See Lucy McRae, "Compression Cradle," 2020, www.lucymcrae.net/compression-cradle. See also, for example, Charles E. Binkley, Michael S. Politz, and Brian P. Green, "Who, If Not the FDA, Should Regulate Implantable Brain-Computer Interface Devices?," *American Medical Association Journal of Ethics* 23, no. 9 (September 2021): 745–49, journalofethics.ama-assn.org/article/who-if-not-fda-should-regulate-implantable-brain-computer-interface-devices/2021-09.

[25] John Paul II, *Man and Woman He Created Them: A Theology of the Body*, trans. by M. Waldstein (Boston: Pauline, 2006), 152.

persons who become sacrament in and through the body.[26] Although the Church does not outright forbid modification of the body (especially in a healthcare context that strives to preserve and heal), it cautions against a certain sense of "morphological freedom" (to use a transhumanist term) that can threaten corporeal integrity, lead to the absolutization of the body, and promote a cult of the body, as it were.[27] In the end, it will be important for us to reflect on the role of technology in replacing bodies or assisting bodies when larger society has chosen to ignore bodies at times.

Anselm Ramelow, OP: Yes, Cory, what Pope John Paul II argues against is a kind of Cartesian dualism. Indeed, our having a body is not like having a car… or a computer, for that matter. But unlike a merely corporeal object, we relate to and have our body. As a consequence, we are not just moved by other objects, but we lead our lives.

Also, I wonder if in emphasizing the embodied aspect of our nature, we are underrating our human distinctiveness from animals. Should we not also defend humans against AI on the basis of human spirituality? Bodies are material, and if anything, computers are material entities—and only that. Just focusing on embodiedness will not make that distinction. It may even reinforce the contemporary "cult of the body" that you mention; and it also leaves angels without their proper status! The importance of the human body (reinforced in the incarnation and the sacraments) has to do specifically with a body that is spiritually animated. What can be done to better spell this out?

Noreen Herzfeld: Jeffrey Pugh has suggested that our fascination with AI and transhumanist goals that consider either uploading our minds to computers or making intelligent computers our progeny represents a return to a Manichaean form of Gnosticism that views the material world as evil and the spiritual/intellectual world as good.[28] I certainly do get the sense in reading works by folks like Kurzweil that they think the body is something to be gotten rid of and our identity is coterminous with our brain. This is contradicted by recent work by neuroscientists such as Antonio Damasio who writes that while "any theory that bypasses the nervous system in order to account for the existence of minds and consciousness is destined to failure ... any theory that relies exclusively on the nervous system to account for minds and consciousness is also bound to fail."[29]

---

[26] Kathleen A. Kalb, "'Theology of the Body' Underpins Health Care," *Health Progress* 93, no. 2 (March-April 2012): 43.
[27] *Catechism of the Catholic Church*, no. 2289.
[28] Jeffrey Pugh, "The Disappearing Human: Gnostic Dreams in a Transhumanist World," in *Religion and the New Technologies*, ed. Noreen Herzfeld (Basel: MDPI, 2017), 51–60.
[29] Antonio Damasio, *Feeling and Knowing* (New York: Penguin Random House, 2021), 21.

Cory Labrecque: Some Christian ecotheologians, like Sallie McFague, drawing upon the incarnation and the sacraments which give certain value to the physical world, will say that the resurrected Christ "is present in and to *all* bodies" and that, ultimately, "*all* bodies can serve as ways to God."[30] Being a "body among bodies" emphasizes, at least to some degree, an important commonality and solidarity in our creatureliness (after all, humans and animals alike were made from the dust of the ground, Gen 2:7, 19). There is a deep sense of interrelatedness and interdependence among bodies that cannot, and should not, be cast aside here.

All of this said, it is the human person—a body-soul composite, whose spiritual dimension ought to be understood together with the physical, social, and historical—who alone is created in/as the *imago Dei* for relationship.[31] Here the distinction between the human-as-body and the non-human-animal-as body (or other bodies for that matter) is made plain, I think.

Moderators: Moving from body to mind, how might AI technology enlighten (or complicate) theological and philosophical perspectives on the meaning of intelligence and consciousness?

Andrea Vicini, SJ: One wonders whether "intelligence" is the most appropriate term to describe algorithmic computation and analysis. The term "artificial intelligence" is so commonly and widely used that it is pointless to even consider proposing to replace it. Still, I would prefer to reserve "intelligence" for the unique and, until now, unmatched abilities of human intelligence, with all its strengths and limitations.

Jordan Joseph Wales: This is an important point to explore. "Artificial intelligence" began as a reflection of mid-century self-understandings; now—for ill—it sometimes is taken more as a defining point of reference than as a reflection.

A thousand years ago, *intellectus* meant the intuitive grasp of something as it is in itself; *intellectus* was the clear vision underlying all discursive reasoning.[32] In the 1950s, AI meant the computational accomplishment of feats that would ordinarily require human thinking and insight: planning, chess-playing, etc. In the 1980s, as robotics became more popular, the logicist reduction of *intellectus* to

---

[30] Sallie McFague, "The Scope of the Body: The Cosmic Christ," in *This Sacred Earth: Religion, Nature, Environment*, 2nd ed., ed. Roger S. Gottlieb (New York: Routledge, 2004), at 262 and 266.

[31] International Theological Commission, "Communion and Stewardship: Human Persons Created in the Image of God," 2004, 1.9–10, www.vatican.va/roman_curia/congregations/cfaith/cti_documents/rc_con_cfaith_doc_20040723_communion-stewardship_en.html.

[32] Josef Pieper, *Leisure: The Basis of Culture* (San Francisco: Ignatius, 2009).

computation was followed by a further reduction of the AI to a "rational agent" that "acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome."[33] Here, then, with historian Yuval Noah Harari, we may re-describe "intelligence" as "the ability to solve problems."[34] Projected onto humans, this approach reduces us to the "instrumentalized reasoning" that Charles Taylor and Alasdair MacIntyre identify as characteristic of our age. If our intelligent machines are intelligent in behaving so as to fulfill our purposes, then are our neighbors also intelligent insofar as they conform to our purposes? Under such a view, Taylor writes, all things are "open to being treated as raw materials or instruments for our projects."[35]

This, of course, is the pride that Augustine considers to be the root and deepest outcome of the fall. The reduction of intelligence to logic, and then to behavior—without reference to an interior life—risks shifting our cultural language so as to depict human life as a task of optimizing (my) benefit, to the exclusion of mutual self-gift. At the limit, we may come to see one another (and even ourselves) simply as behavior-producers, whose value will be quantifiable in terms of the production of desired actions. With the recent rise in the tracking of personal activities, habits, fitness, and performance—despite the obvious benefits of these technologies—we may see this shift already in progress.

Paul Scherz: I really like how Jordan provides a historical outline of the understanding of intelligence in AI. However, I wonder if we have not moved on to a fourth stage beyond the movement from classical *intellectus*, to computation, to instrumental reason. With contemporary forms of machine learning, as they are being deployed across the economy and government, the goal seems merely to make predictions, things like the behavioral futures that Shoshanna Zuboff discusses.[36] Intelligence becomes something akin to gambling skill.

What I find interesting is how models of intelligence used in the programming realm feed back into areas of human activity. As I and others have noted, the widespread use of machine learning and Big Data is transforming many fields of science such as genetics, which is

---

[33] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. (Upper Saddle River, NJ: Pearson, 2009), 2.

[34] David Kaufman and Yuval Noah Harari, "Watch Out Workers, Algorithms Are Coming to Replace You—Maybe," *New York Times*, October 18, 2018, www.nytimes.com/2018/10/18/business/q-and-a-yuval-harari.html.

[35] Charles Taylor, *The Ethics of Authenticity* (Cambridge, MA: Harvard University Press, 1991), 5.

[36] Shoshanna Zuboff, *The Age of Surveillance Capitalism* (New York: PublicAffairs, 2019).

flooded with genomic data.[37] There comes to be an expectation that scientific knowledge and discovery will come merely from having machines churn through ever larger piles of data, with some even suggesting that AI systems will be able to perform their own research in an era of "hypothesis-free" research. The ways these changed concepts of intelligence and understanding filter back into scientific practice are causing significant distortions in contemporary research. I would imagine that these kinds of effects are being seen in a number of fields.

Brian Cutter: Very interesting thoughts, Jordan. And, like Paul earlier quoted, I am reminded of one of my favorite passages from C. S. Lewis's *The Abolition of Man*:

> If man chooses to treat himself as raw material, raw material he will be: not raw material to be manipulated, as he fondly imagined, by himself, but by mere appetite, that is, mere Nature, in the person of his dehumanized Conditioners. ... Either we are rational spirit obliged for ever to obey the absolute values of the *Tao* [natural law], or else we are mere nature to be kneaded and cut into new shapes for the pleasures of masters who must, by hypothesis, have no motive but their own "natural" impulses.[38]

Jordan Joseph Wales: Thank you, Brian. From a similar time period as Lewis, we might also cite Winston Churchill, who believed that not the British but the Soviet society would be best suited for robotic slaves because they would be the final fulfillment of what Churchill saw as the Soviet view of the person as a cog in the machinery of state. But now we find the same view attributable to tendencies in our own society, as Lewis foresaw.[39]

Marga Vega: Jordan, you make an important point about the history of AI research. Under the computational theory of the mind and cognitivism, the first years of artificial intelligence encouraged computer scientists' hope to achieve machines that could think not just *like* humans but also *better than* humans, possibly even showing consciousness. Conversely, it also opened the prospect of mapping the human mind in computational terms, dismissing the importance of consciousness and awareness in cognition, and leveling any assumed

---

[37] Paul Scherz, "The Displacement of Human Judgment in Science: The Problems of Biomedical Research in an Age of Big Data," *Social Research* 86, no. 4 (2019): 957–76; Erik Larsen, *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do* (Cambridge, MA: Belknap, 2021); Jenny Reardon, *The Postgenomic Condition* (Chicago: University of Chicago Press, 2017); and Hallam Stevens, *Life Out of Sequence* (Chicago: University of Chicago Press, 2013).

[38] Lewis, *The Abolition of Man*, 73.

[39] See Churchill's essay "Fifty Years Hence," in *Thoughts and Adventures: Churchill Reflects on Spies, Cartoons, Flying, and the Future*, ed. James W. Muller, Paul H. Courtenay, and Alana L. Barton (Wilmington, DE: ISI, 2009).

ontological differences between the human mind and machine intelligence.

Underlying both projects is the analogy that the mind is to software as the brain is to hardware. How far we take this analogical seesaw by conceding more weight to the idea that machines have minds or to the idea that human minds are machines may not matter much if the result in both cases is to minimize the ontological differences between minds and machines under a paradigm that equates computation with intelligence.

The problem with equating computation and intelligence is that computation is only possible if there are minds relative to which we can assign computational interpretations. In other words, computation cannot ground intelligence because intelligence is an *a priori* condition for computation. John Searle's Chinese Room Argument (CRA), which initially pointed at the lack of semantics in computers, later addressed this difficulty with the thesis "syntax is not physics."[40] In comparing minds and computers, the CRA noted that it is not only that computers have a syntax and not semantics; they do not even have a syntax since any syntactical structure is observer relative. Syntax exists only relative to minds capable of mental content, and that is precisely what is at stake in the case of computers: the capacity to have something other than purely physical causal processes devoid of mental content.

Jordan Joseph Wales: Marga, I like the way you are going here. Even before we speak of a soul, we must ask whether the chemical reactions in the nervous system have some causality beyond that which is describable in physics and chemistry. If physics and chemistry as presently understood exhaustively describe our bodily processes, then there is neither consciousness nor meaning—a claim that seems manifestly false by our very experience.

Anselm Ramelow, OP: Indeed. When we talk about ourselves, what we mean by "consciousness" has features we do not expect machines to have, among them a subjectively experienced point of view, intentionality and, for rational minds, a kind of reflexivity that cannot be instantiated in material objects.[41] Another feature is a certain

[40] Searle introduced the idea of syntax being observer-relative in his Presidential Address to the American Philosophical Association, and it has appeared since in subsequent formulations of the CRA such as "Who is Computing with the Brain," *Behavioral and Brain Sciences* 13, no. 4 (December 1990): 632–42, and *The Mystery of Consciousness* (London: Granta, 1997). Sometime after 2003, the argument appears as the "syntax is not physics" thesis in Searle's lecturing and writings.

[41] A point also made by Karl Rahner, SJ, "Person. II. Man. C: Theological," in *Sacramentum Mundi: An Encyclopedia of Theology*, vol. 4: Matter to Phenomenology, ed. K. Rahner (Montreal: Palm, 1969), 417. Reflexivity is also the root of creativity; see Anselm Ramelow, OP, "Can Computers Create?," *Evangelization and Culture* 1 (2019): 39–46.

"unified" character: consciousness is a unifier of all its contents. Ontologically, this feature corresponds to the unified life of organisms, which differentiate themselves from their environment both in their actions and in their very being. Characteristically, these are kinds of unities we cannot make ourselves. Living beings originate by procreation, not artificially (*omne vivum ex vivo*). Why would we expect this to be different in the case of consciousness, which is a life that has the additional unifier of awareness? The making of conscious entities may require, therefore, the causality of someone who gives things both their nature and existence, the most fundamental unifying properties. Such a maker would therefore need to be a creator (God). We, on the other hand, *presuppose* the existence of things and *rely on* their natures in order to build artifacts with them. These artifacts do not have any other unity than the purpose we have for them. The unity is not ontological or intrinsic to them, but only in our minds. This is true for AI as well: neither in its being nor in its operations does AI have the requisite unity to be conscious. Metaphysically, the parts are in potency with regard to the whole; hence the actualization of this unified whole requires a proportionate cause. If the unity in question concerns the very nature of the thing, this cause may need to be a creator.

Marga Vega: That is a relevant question, Fr. Anselm, whether consciousness can pertain to entities that are not alive, and whether consciousness is itself a type of life. If the latter happens to be the case, it seems that a conscious artifact is not possible. However, some would defend the proposition that perhaps intelligent computers do not need consciousness—all that is required is intelligent behavior. It is questionable that what is meant by "intelligence" in the case of humans and in that of computers can be taken univocally.

But even if we generously granted "intelligence" to computers, their status as artifacts and non-persons would remain. Even for those unfamiliar with Boethius's definition of the person as an individual substance of rational nature, the idea that rationality grounds our personhood takes hold of our minds both through our civilization's history and our personal and societal values. Based on this intuition, some have questioned, with perplexity, the personhood of human beings whose rationality is impaired. Embryos, neonates, people in vegetative state, or those with disabilities may lack the exercise of intellectual capacities that some would consider essential for personhood.

Likewise, based on intelligence, a debate emerges on whether machines could have, if not metaphysically then at least legally, the status of persons. If we have machines that compete with us in terms of intelligence, should they also qualify as persons if intelligence characterizes personhood? The paradox is that placing intelligence as the paragon for personhood may strip the title of "person" from humans with dormant rational capacities while entertaining whether machines could be eligible candidates for this status. The challenge of AI offers

us an ontological opportunity: perhaps intelligence is not a definitive measure for personhood, if personhood (or even humanity) does not ensue from the possession of an ability. On the contrary, personhood precedes any capacity.

Jordan Joseph Wales: Marga, I understand wanting to uphold the personhood even of those who have dormant faculties, but you make it seem as if denying personhood to machines is a foregone conclusion.

Marga Vega: We tend to infer what something is from the way it acts. At first sight, it would seem that: (1) if a computer acts intelligently, then it is intelligent; and therefore (2) it can be counted as a person. From the point of view of how we get to know things, this would seem like a valid inference. But we must not confuse epistemology with ontology, how things are.

First, behavior alone does not guarantee that what causes the behavior is the same in both cases. A sore throat may be a sign of the flu but also of COVID-19. Performing specific intelligent tasks may have a comparable output by a computer and a human, but the causal elements could be very different. Therefore, we cannot conclude intelligence from the appearance of intelligent behavior: we need independent definitions of what counts as intelligence and what kind of causality it requires.

Second, it is questionable whether intelligence or rationality constitutes persons (granted that rationality accompanies personhood). It could well be that rationality does not ensure personhood and that personhood causes rationality. In this case, we would have things backward in assigning personhood to computers based on their intelligence. Therefore, we would need to inquire into what is the root of personhood in the first place.

Cory Labrecque: I think Marga raises an important point here that brings to the fore contemporary wrestling with the definition of personhood. The concept is at the center of bioethical discourse, but so few agree on how it should be understood.

In a short piece entitled "Is Koko a Person?" James W. Walters—Professor of Ethics at Loma Linda University—makes a distinction (well known by theorists who study moral status) between what he calls physicalism and personalism (not to be confused with other uses of this term in philosophy and theology). The former argues that "the essence of a person is found in his or her biological make-up. All humans are persons, ipso facto." The latter, which is telling here and links to Marga's critiques, locates the essence of a person "in one's mental capacities and ability to use these in satisfying ways. Whether one is a human is not important." In this way, robots and computers could fit the bill, while certain human beings—bereft of certain functions—fall short. This linking of capacity, function, and performance

to dignity and value further deprecates other shared dimensions of humanhood: fragility and vulnerability.[42]

Noreen Herzfeld: I think it is worth noting that "person" has become a legal category here in the US. We allow corporations to be classed as persons, in this legal sense. One problem with "personhood" is that it is a binary—one either is or is not a person. In legal terms, it must be binary, however this makes it less than useful as a philosophical designation. With respect to AI, the fetus, or the severely disabled, I think we would do better to speak in shades of gray, rather than black or white.

Brian Cutter: While I suspect that AI technology will not teach us much about the nature of consciousness, I do want to say there is a lot here we probably cannot really know. If we eventually create an AI that passes behavioral tests for general intelligence (e.g., a Turing test), we probably will not know whether it is conscious, even if it says it is.

In my view, consciousness (i.e., subjective experience) is ontologically distinct from any set of physical or computational processes, so even if we had complete knowledge of the machine's physical operations, this would not conclusively settle whether it was conscious.[43] While consciousness is distinct from any purely physical process, conscious states are obviously *correlated* with certain physical processes (e.g., processes in human brains) in regular, lawful ways. To figure out whether an advanced AI would be conscious, a key question is whether the "psychophysical laws" (the laws of nature by which physical states are linked to states of consciousness) are *substrate-independent*—that is, whether they are sensitive to the material composition of a physical system, or whether they are only sensitive to the higher-level causal organization of the system, abstracting away from its material substrate.

In principle, the high-level causal organization of a human brain could be implemented in a computer. For example, a detailed computer simulation of a human brain would exhibit the same causal organization as a human brain, but it would be realized in a silicon-based material substrate rather than a carbon-based substrate. If the psychophysical laws are substrate-independent, as some philosophers have argued, then a detailed computer simulation of a human brain would

---

[42] See James W. Walters, "Is Koko a Person?," *Dialogue* 9, no. 2 (1997), circle.adventist.org/files/CD2008/CD2/dialogue/articles/09_2_walters_e.htm.

[43] I will not defend the ontological distinctness claim here; I accept it on the basis of arguments like those given in David J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (Oxford: Oxford University Press, 1996); Brian Cutter, "The Modal Argument Improved," *Analysis* 80, no. 4 (2021): 629–39; Saul A. Kripke, *Naming and Necessity* (Cambridge, MA: Harvard University Press, 1980); and Adam Pautz, "Do Theories of Consciousness Rest on a Mistake?," *Philosophical Issues* 20, no. 1 (2010): 333–67.

be conscious.[44] My own view is that we currently have no idea whether the psychophysical laws are substrate-independent, and I doubt this question will be settled any time soon.

Andrea Vicini, SJ: As a final comment on this question, I would just like to point out that concepts such as intelligence, embodiment, natural/artificial, and consciousness do not exist in a vacuum. Each of these concepts tries to articulate a particular dimension of the complex human and social reality with its plurality.[45] Moreover, how one understands these concepts depends on the historical and cultural contexts in which they are articulated. Critical reasoning should help us examine the elements that characterize our context and how this context influences our understanding of each one of these concepts.[46] Asking whether or not they promote the common good could help to discern between the various interpretations of these multiple concepts and the concrete implementations they make possible. At the moment, sadly, AI seems to reinforce the social inequities, discriminations, and biases present in our society.

Moderators: Fr. Andrea, this is a good segue to our next point. For each of you, what are some key ethical issues to focus upon with respect to AI? How might Christians and the Catholic Church in particular helpfully respond to these issues?

Noreen Herzfeld: Well, as a first point, I would refer to my article in this issue on lethal autonomous weapons systems (LAWS) and just war theory.[47] We need to push for bans on LAWS before they become widespread. Should these weapons ever be designed to act with complete autonomy (no human in the loop) the likelihood of unforeseen consequences would be staggeringly high, as would the likelihood of these weapons being deployed by rogue actors. While no ban is totally enforceable, international condemnation does have an effect, as we have seen with chemical weapons. I find it both interesting and heartening that most military and former military generals I have spoken to

[44] See, e.g., Chalmers, *The Conscious Mind*.
[45] See Peter G. Kirchschlaeger, "Artificial Intelligence and the Complexity of Ethics," *Asian Horizons* 14, no. 3 (2020): 587–600.
[46] See Paolo Benanti, "Algor-éthique: intelligence artificielle et réflexion ethique," *Revue d'ethique et de théologie morale* 3, no. 307 (2020): 93–110. See also Paolo Benanti, *Digital Age: Teoria del Cambio d'Epoca: Persona, Famiglia e Società* (Cinisello Balsamo: San Paolo, 2020); Paolo Benanti, *Realtà Sintetica: Dall'Aspirina alla Vita: Come Ricreare il Mondo?* (Roma: Castelvecchi, 2018); Paolo Benanti, *Le Macchine Sapienti* (Bologna: Marietti, 2018); Paolo Benanti, *Oracoli: Tra Algoretica e Algocrazia* (Roma: Luca Sossella, 2018).
[47] Noreen Herzfeld, "Can Lethal Autonomous Weapons Be Just?," *Journal of Moral Theology* 11, Special Issue 1 (2022): 70–86.

are adamantly against the design or deployment of fully autonomous weapons.[48]

Levi Checketts: Another pressing problem is the displacement of laborers and the increase of wealth disparity across the globe. One of the most immediate practical uses of AI involves cost-saving and labor-saving procedures. The ultimate result of this will be that those who own or control AI will become fabulously wealthy while the vast majority of others will find themselves competing against a labor system that does not need housing, time off, or provision for biological necessities. Magisterial Catholic social teaching, such as *Laborem Exercens* or *Pacem in Terris*, reminds us that work is a human good and governments have an obligation to ensure the common good above the amassing of wealth (*Laborem Exercens*, no. 17; *Pacem in Terris*, nos. 56, 121). Here, the Church has a vast treasure of resources to turn to, prophetic voices condemning the unthinking use of power to the benefit of few and detriment of many. We might think of the witnesses of liberation theologians, St. Ambrose of Milan or John Chrysostom, and speak out against the unjust use of power and wealth against the poor.[49] AI should be used for *all* humanity, not only the rich.

Paul Scherz: Levi, I agree that the way AI applications support concentrations of economic and political power is a huge problem. These applications require immense datasets and computing power, so they can be deployed only by large corporations, governments, or other entities with the requisite funding and access to those resources. In workplace settings, AI can be implemented in ways that support deskilling and the centralization of knowledge in management, thus continuing the trend of worker disempowerment seen in Taylorism.[50] Concentration of power has long been considered a problem in Catholic social thought, insofar as it increases social struggles and decreases the possibility for free action.[51] We see this anew in the way these forms of concentrated power can disempower workers, undermine privacy, and expand bias.

---

[48] See Noreen Herzfeld and Robert H. Latiff, "Can Lethal Autonomous Weapons be Just?," *Peace Review* 33, no. 2 (2021): 213–19.

[49] See, for example, John Chrysostom, *Homily 7 on Colossians*; Thomas Aquinas, *Summa theologiae*, IIa IIae, q. 66, a. 1; and Gustavo Gutiérrez, *A Theology of Liberation: History, Politics and Salvation*, trans. Caridad Inda and John Eagleson (Maryknoll, NY: Orbis, 1973), 163–64.

[50] For Taylorism and a more general account of the degrading effects of routinization on workers, see Harry Braverman, *Labor and Monopoly Capitalism: The Degradation of Work in the Twentieth Century* (New York: Monthly Review, 1998). For broad accounts of deskilling in the wake of automation, see Nicholas Carr, *The Glass Cage* (New York: Norton, 2014); and Shannon Vallor, "Moral Deskilling and Upskilling in a New Machine Age," *Philosophy and Technology* 28, no. 1 (2015): 107–24.

[51] E.g., Pius XI, *Quadragesimo anno: On the Reconstruction of the Social Order* (1931), nos. 105–109, www.vatican.va/content/pius-xi/en/encyclicals/documents/hf_p-xi_enc_19310515_quadragesimo-anno.html.

Andrea Vicini, SJ: Agreeing with Levi and Paul, threats to social justice, including racial discrimination and increasing inequities, are relevant ethical challenges.[52] Technological progress fostered by AI ought to promote greater equality by addressing the increasing gap between those who have and those at the margins of the social fabric. AI should not further heighten social inequities. Christians and Catholics, both as individuals and ecclesial institutions, share social responsibility towards fostering awareness on the part of citizens and believers regarding uses of AI technology that disempower and marginalize, and to join multiple social actors (e.g., citizens, groups, and organizations—nationally and internationally) in addressing these diverse ethical challenges in collaborative ways with scientists, politicians, activists, communities, and multinational companies. Finally, with a variety of its agencies and the leadership of Pope Francis, the Vatican appears to be at the forefront of dialogue, reflection, and critical engagement regarding AI involving scientists, scholars in the humanities, universities, and biotech companies.[53] Such an engagement is praiseworthy and shows how it is possible, even necessary, to be participants in the social arena by joining multiple social forces while aiming at promoting a broad social agenda, open toward progress and the future, and animated by a realistic hope.

Cory Labrecque: The issues are myriad. One topic that does not often come to the fore is the importance of touch for healing, and how the transfer of care (or even parts of the care process) to AI software or machines may very well suppress crucial bodily elements of the patient-healthcare provider relationship (that is, of human bodies in relationship) conducive to well-being. In her *Broken Nature* exhibit, sci-fi artist Lucy McRae introduces a new work that she calls a "compression cradle" as a response to the "touch crisis" in which we find ourselves due to the lack of physical contact that has come about from our excessive connection with technology. Yet McRae responds to this mark of rampant technologization by introducing yet another technology: a machine that "affectionately" squeezes the body through a series of aerated membranes that "hold you tight in an attempt to prepare the self for a future that assumes a lack of human touch."[54]

As another point, for the Church, technology must have the good of human beings and the whole human family at its heart. It must be an expression of stewardship and service, contribute to genuine progress (that is, a progress that will lead human beings to exercise a

[52] As an example, see Isabel Wilkerson, *Caste: The Origins of Our Discontent* (New York: Random House, 2020).

[53] See Vincenzo Paglia and Renzo Pegoraro, *The "Good" Algorithm? Artificial Intelligence: Ethics, Law, Health,* Proceedings of the XXVI General Assembly of the Pontifical Academy for Life (Rome: Pliniana, 2021).

[54] See Lucy McRae, *Compression Cradle* (2020), www.lucymcrae.net/compression-cradle.

wider solidarity and opening themselves more freely to others and to God) (*Octogesima Adveniens*, no. 41), respect the inherent dignity of human beings and all natural environments, and recognize the delicate complexity of ecosystems and the interdependencies extant within them. Although technology may very well extend our dominion over the material world (*Caritas in Veritate,* no. 69), the Church reminds us that the Biblical mandate to subdue the earth and have dominion over it is an *entrusted* one that should never become despotism or absolute mastery/lordship over the body (one's own or others' bodies). The mandate is very much a collective responsibility to make manifest God's love for the whole of Creation. The Church—for whom the corporal works of mercy (i.e., feeding the hungry, tending to the ill, clothing the naked, sheltering the homeless, etc.) shape the Christian moral life as an extension of God's compassion—ought to be on the front line countering these trends.

Marga Vega: Going in a different direction, I think there is a real risk in the delegation of moral decision making to artifacts. Rationality is a capacity that only makes sense as an ability *from* and *for* the ability's owner. Rationality is not an absolute and cannot be uprooted from a teleology nor cut off from someone who holds that rationality. The unique challenge AI poses is what to do when the tool becomes the wielder. Might the wielder then become a tool in the service of the new wielder: a machine? Our created AI could become so self-sustaining and independent that it could hand out decisions leaving us in the dark as to the *criteria* guiding the reasoning process and powerless to resolve what is best. We do not need to jump into a self-driving car to envision scenarios where the bliss of ignorance and referred decision-making can become a liability. It is clear then that we cannot assemble an artificial intelligence without built-in values that guide and preserve personal rational criteria. More than ever, ethics is necessary for technology, not just as *how to decide the proper use* of the technological invention, but *how to build-in values* in the very fabric of the tool's constitution.

Jordan Joseph Wales: Taking the artificial agent question in a different direction again, I have already described the quandary of ethical formation that will arise from our owning the services of apparent but unreal persons. Several of us have also attended to how our beliefs concerning AI may reshape or distort our understanding of the human person. Therefore, Christians and the Church must above all bear witness to the importance of the human interior life and to the self-gift that flows from it. Even before and beyond Christians' and the Church's declarations concerning AI, it is this relational witness that will preserve in our culture the means by which to live humanly alongside and with the fantastic technological developments of today and tomorrow.

Noreen Herzfeld: Building on that, one place where self-gift is most evident is in our sexual lives. In the act of intercourse, we give our bodies to another in an openness and vulnerability that ideally flows from the interior life Jordan speaks of. Increasingly lifelike robots, or sexbots, are already being developed to function as our electronic lovers. While these might make interesting or even desirable sexual partners, they represent another form of idolatry, substituting a relationship with the living with something made, and thus controlled, by our own hands. In this way we risk reducing sex to a one-way street, in which the robot is there to meet our needs and proclivities, emptying the act of its wildness and mystery and making few demands upon us. It becomes a form of whoredom.[55]

Brian Cutter: For me, the most philosophically interesting ethical question about AI is whether an advanced AI would itself have moral status—whether it would have morally significant interests we ought to respect (my concern here is with hypothetical future AI, not current AI). This would partly depend on whether it is conscious (i.e., capable of subjective experience). A capacity for subjective experiences like pleasure and pain is, I think, a sufficient condition for having *some* moral status, though not a sufficient condition for the *full* moral status associated with persons. Thus, even if an advanced AI with the right cognitive architecture would be conscious, and therefore have some moral status, it might not have full moral status. We should also think about how to navigate the issue of *moral risk* in this domain. How should we treat an advanced AI if we are unsure whether it has any moral status (say, because we are unsure whether it is conscious?). It would be interesting to explore the analogies and disanalogies to the issue of moral risk in debates about the ethics of abortion. A common pro-life argument is that abortion is gravely wrong even if we are unsure whether a fetus is a person, since in general, it is wrong to do things that carry a significant risk of killing an innocent person.

Anselm Ramelow, OP: Brian, the "moral risk" question is interesting. Still, there are two disanalogies here: (1) in the one case there is a high-performance entity where we are unsure of consciousness, in the other we do assume neither consciousness nor expect high performance yet think mistreatment risky. Why? (2) The causal history of both entities is different: one we have made, the other begotten. Life comes only from life, should consciousness not come a fortiori from conscious beings? Maybe that answers also point (1)?

Brian Cutter: Fr. Anselm - Interesting points! I agree there is some disanalogy. Most importantly, the "risk" is much greater in the abortion case, since the fetus is a member of the human species, and it

[55] See Noreen Herzfeld, "Religious Perspectives on Sex with Robots," in *Robot Sex: Social and Ethical Implications*, ed. John Danaher and Neil McArthur (Cambridge, MA: MIT Press, 2017), 91–102.

is extremely plausible that being a member of the human species is a sufficient condition for full moral status. Pro-choice philosophers typically reject the latter claim on the grounds that the fetus, though a human being, does not have sufficiently developed mental capacities to qualify as a "person." That objection has implausible and repugnant implications, e.g., infants and the severely mentally disabled lack full moral status. I do not think the principle that conscious beings only come from conscious beings settles the question. AI comes from us, and we are conscious. Here I assume "comes from" covers creating and not just begetting; otherwise, the principle that "life only comes from life" is false, since the first life forms were not begotten by any living thing.

Anselm Ramelow, OP: Yes, indeed. Here, further arguments would need to be made that the first making of life can only be done by a Creator. We at least cannot; and to me that seems to be an important difference.

There is an opportunity for Catholic philosophy and theology to speak on issues related to AI. Contemporary thought is hesitant to reflect on the whole of reality. Metaphysical questions or questions of the meaning of life do not typically receive robust answers. The Catholic Church is, for theological reasons, more confident in the ability of our mind to propose such answers, and she has a long history of formulating such answers. It is time to take such proposals from the shelf and articulate them anew. This need not require an attempt to prove the truth of these proposals. It may rather be proposed as an "inference to the best explanatory hypothesis." If it gives a richer and more cogent explanation of reality, including of humans and AI, and accounts for more data, including for our moral intuitions, then it has the chance to be helpful as a response.

Moderators: Fr. Anselm, that is a great transition to our last question about the practical relevance of these questions about AI. What can and should Catholic institutions such as universities, hospitals, charities, and the Vatican, and Christian institutions more broadly, do in order to facilitate the better uses and restrict the worse uses of AI?

Jordan Joseph Wales: Catholic institutions must become well educated as to how AI works and reflect deeply both on AI and the human person in order. So doing, both individuals and institutions can, as members of the ecclesial body, advocate appropriately, for instance, for laws that prohibit the mistreatment of apparent persons (on the basis not of personal rights, which they cannot have, but of their signifying of the personal, much like public anthropomorphic works of art). Persons of any faith or none should remind themselves again and again, that the instrumental functioning within which we use these

tools does not exhaust the meaning of the human person. However, that is not all that we must say; we must consider also how the theology of creation and the human person allows us to think more carefully about just *what* this or that AI might be. The theory of biological evolution invites theological reflection on an unfolding divine providence in light of ancient Christian beliefs concerning the manifestation of God's wisdom in the created order. So too, the rise of artificial intelligence—especially deep learning and its capacity to apprehend hidden dynamics within large data sets—allows us to think again about the ways in which machines designed for our purposes can be both attuned to the deep dynamics of contingent events (e.g., markets, societies, and the weather) and can also obscure those dynamics (e.g., in AI bias) depending on how we have carved up the reality we seek to engage. Human engagement with the world is, from a Catholic point of view, a theological and a spiritual phenomenon; the more Catholic institutions and theologians reflect on artificial intelligence, the more—and the more usefully—we shall find we have something to say.

Andrea Vicini, SJ: Institutions of higher education play an important social role with their teaching and research. They contribute to the education, formation, and training of students and citizens in engaging the technological developments and social implementations of AI in critical ways, in light of an articulated ethical approach, with a strong attention given to social dynamics and their historical implementation.[56] For example, the history of medicine, technology, and science allows us to learn both from virtuous and vicious approaches by considering, respectively, benefits and troubling consequences and addressing any injustice and inequity.

Paul Scherz: I agree with Andrea's point as to the institutional importance of Catholic colleges and universities. They can act as both important research centers for exploring these questions as well as centers for forming the next generation of citizens in using these technologies well. In the latter role, Catholic institutions of higher education can serve as a crucial witness as to how to embody the use of the technologies well, or as a prophetic witness as to what instances and uses of these technologies must be rejected.

Catholic health care is even closer to the front lines on these issues. These health systems have vast troves of data on their patients, and technology companies are eager to gain access through partnerships that will allow them to sift through the data with their machine learning systems. These health systems must be careful about, on the one hand, falling to the hype about the promise of artificial intelligence and thus overpromising what these partnerships might achieve and, on the other hand, using their patients' data in exploitational ways. The

[56] See Angelo Chakkanattu, CMI, "Artificial Intelligence: Human Natural Machine Intelligence of Evolution," *Asian Horizons* 14, no. 3 (2020): 563–86.

structure and implementation of such partnerships will help determine whether these programs respect important goods such as dignity, privacy, the common good, and service to the poor, or whether they make use of the data solely for instrumental goods of profit or, worse, a biased delivery of healthcare. Even with the best intentions, a poorly structured program could lead to dangerous practical effects. This is a place where moral theologians along with scholars from other fields such as law could, in a very practical manner, help these systems fulfill their role as a ministry of the Church. Scholars can assist in examining how to prevent the dangers and encourage the positive potential of these partnerships. Healthcare is another area in which Catholic institutions could be a witness as to how to use these technologies well.

Levi Checketts: One risk I see is people rushing headlong into the technology, seemingly pursuing what Max More calls the "proactionary principle"—focusing on developing technology first and adjusting it to the good later.[57] I have been party to many conversations where technologists aptly demonstrate their knowledge of the field of AI and how *they* think it might be good, but they often think their opinion of the moral problems surpasses the understanding of theologians and philosophers. One reason why this problem exists, in my view, is that theologians involved in interdisciplinary discussion cut right to engineers and entrepreneurs as dialogue partners, rather than dialoguing with technology scholars, policy advisors, social theorists, philosophers of technology, and critical theorists. We need to have more events to discuss theology and AI, but we need to open the forum so that scholars, activists, and ministers who are not inherently engaged with AI production are discussing it. In my research on transhumanism, with dozens of responses from theological thinkers, one of the single best responses I read was that of James Keenan, who has no interest whatsoever in transhumanism. As an outside scholar, he was able to articulate problems that many too close to the issue had missed, such as the nature of Catholic collectivism and embodiment. Likewise, if we invite feminist theologians, ecological theologians, critical theorists, black, Latinx, Asian theologians, and others into the dialogue, we may find creative responses and ideas for the problems at hand.[58]

Noreen Herzfeld: Levi makes a good point. We need to remember that most AI research is funded with soft money. This means the researchers must hype the possible good outcomes that will come out of it. As one technologist at MIT put it "We shall overclaim!" Yet we cannot have all the critique coming from the outside. Many of the ethical dilemmas that appear in our technology are baked into the design. Thus, it is imperative that designers themselves start asking, not only

[57] Max More, "The Proactionary Principle: Optimizing Technological Outcomes," in *The Transhumanist Reader*, 258–67.
[58] Keenan, "Roman Catholic Christianity—Embodiment and Relationality, 155–72.

what good their product can do, but also what harm. When harm occurs, the corporations that design, market, and run our computer systems need to take responsibility. Ultimately, a computer, as a non-sentient thing, remains a tool. It cannot be a moral agent. Only humans are that.

Moderators: Thank you, Noreen; on that note we have run out of time. This has been a great conversation! Thank you to all of you for your contributions.

Brian Patrick Green is Director of Technology Ethics at the Markkula Center for Applied Ethics at Santa Clara University; Matthew J. Gaudet is Lecturer of Engineering Ethics at Santa Clara University and Fellow at the Grefenstette Center for Ethics in Science, Technology, and the Law at Duquesne University; Levi Checketts is Assistant Professor of Religion and Philosophy at Hong Kong Baptist University; Brian Cutter is Associate Professor of Philosophy at the University of Notre Dame; Noreen Herzfeld is the Nicholas and Bernice Reuter Professor of Science and Religion at St. John's University and the College of St. Benedict and senior research associate with ZRS Koper; Cory Andrew Labrecque is Associate Professor of Bioethics and Theological Ethics, and the inaugural Chair of Educational Leadership in the Ethics of Life at the Faculty of Theology and Religious Studies at Laval University; Anselm Ramelow, OP, is Professor of Philosophy at the Dominican School of Philosophy and Theology; Paul Scherz is Associate Professor of Moral Theology and Ethics at the Catholic University of America; Marga Vega is Professor of Philosophy at the Dominican School of Philosophy and Theology; Andrea Vicini, SJ, is Michael P. Walsh Professor of Bioethics and Professor of Theological Ethics in the Boston College Theology Department; and Jordan Joseph Wales is Associate Professor and the John and Helen Kuczmarski Chair in Theology at Hillsdale College.

Journal of Moral Theology, Vol.11, Special Issue 1 (2022): 41–69

Artificial Intelligence and Social Control: Ethical Issues and Theological Resources

Andrea Vicini, SJ

ARTIFICIAL INTELLIGENCE (AI) IS A RAPIDLY expanding field of ongoing technological developments. While many stress how AI is socially beneficial, others manifest their critical assessment by focusing on what is researched and produced, and how it is used. To articulate an ethical analysis that highlights relevant aspects of the social impact of AI, this paper first considers the 2020 joint statement titled *Rome Call for AI Ethics*, which exemplifies an ethical approach centered on principles, as well as recent statements of Pope Francis, which articulate a more comprehensive ethical framework. Second, turning to the social context, the paper focuses on how AI is used within facial recognition systems, the justice system, and workplaces. A brief analysis of social dynamics, structures, and implementation strategies suggests that further ethical resources are needed. Hence, the paper ends with an invitation to discern between an ethic of control and an ethic of risk, engage biopower and biopolitics, and reflect on human labor.

THE *ROME CALL FOR AI ETHICS* AND POPE FRANCIS

On February 28, 2020, at the end of the international workshop "The 'Good' Algorithm? Artificial Intelligence, Ethics, Law, Health," organized by the Vatican's Pontifical Academy for Life (PAL), representatives of the PAL, Microsoft, IBM, the Food and Agriculture Organization of the United Nations (FAO), and the Italian Government signed the document *Rome Call for AI Ethics*,[1] "to support an ethical approach to artificial intelligence and promote a sense of responsibility among organizations, governments, and institutions with the aim to create a future in which digital innovation and technological

[1] Pontifical Academy for Life, "Artificial Intelligence 2020," 2020, www.academyforlife.va/content/pav/en/events/intelligenza-artificiale.html. AI stands for "Artificial Intelligence." For the *Encyclopedia Britannica*, artificial intelligence is "the ability of a computer or computer-controlled robot to perform tasks commonly associated with intelligent beings" (B. Jack Copeland, "Artificial Intelligence," *Encyclopedia Britannica,* 2020, www.britannica.com/technology/artificial-intelligence).

progress serve human genius and creativity and not their gradual replacement."[2]

The *Rome Call for AI Ethics* acknowledges that "AI offers enormous potential when it comes to improving social coexistence and personal well-being, augmenting human capabilities, and enabling or facilitating many tasks that can be carried out more efficiently and effectively."[3] Technology should be developed "for the good of humanity and of the environment, of our common and shared home, and of its human inhabitants, who are inextricably connected."[4]

To advocate for uses of AI technology aimed at serving the "human family,"[5] avoiding any exploitation and "respecting the inherent dignity of each of its members and all natural environments, and taking into account the needs of those who are most vulnerable,"[6] the document relies on the promotion of human rights.[7] Moreover, "the impact of the transformations brought about by AI in society, work, and education"[8] demands the development of "specific curricula that span different disciplines in the humanities, science, and technology."[9]

Finally, six principles summarize the "fundamental elements of good innovation":[10] transparency (i.e., AI systems must be explainable); inclusion ("*the needs of all human beings must be taken into consideration so that everyone can benefit and all individuals can be offered the best possible conditions to express themselves and develop*"); responsibility (concerning both designers and users); impartiality (avoiding bias and safeguarding fairness and human dignity); reliability of the AI systems; security of the AI systems; and respect for the privacy of users.[11]

Principles are highlighted in other international documents. As an example, in June 2019, the G20[12] adopted AI principles that aim at promoting "human-centered" developments and uses of AI

---

[2] Pontifical Academy for Life, "Artificial Intelligence 2020."

[3] Pontifical Academy for Life, "Rome Call for AI Ethics," 2020, https://www.romecall.org/wp-content/uploads/2021/02/AI-Rome-Call-x-firma_DEF_DEF_con-firme_.pdf.

[4] Pontifical Academy for Life, "Rome Call for AI Ethics," 3.

[5] Pontifical Academy for Life, "Rome Call for AI Ethics," 3. The document quotes United Nations, "Universal Declaration of Human Rights," 1948, www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf.

[6] Pontifical Academy for Life, "Rome Call for AI Ethics," 3.

[7] See Pontifical Academy for Life, "Rome Call for AI Ethics," 4–6.

[8] Pontifical Academy for Life, "Rome Call for AI Ethics," 5.

[9] Pontifical Academy for Life, "Rome Call for AI Ethics," 5.

[10] Pontifical Academy for Life, "Rome Call for AI Ethics," 8.

[11] Pontifical Academy for Life, "Rome Call for AI Ethics," 7. Emphasis in the original. In the document, only a few words define each principle.

[12] The G20 is the international forum for the governments and central bank governors from nineteen countries and the European Union.

technology.[13] These principles are: "inclusive growth, sustainable development, and well-being;[14] human-centered values and fairness;[15] transparency and explainability;[16] robustness, security, and safety;[17] and accountability."[18] As the G20 document acknowledges, these principles were formulated in the 2019 *Recommendation of the Council on Artificial Intelligence* of the Organisation for Economic Cooperation and Development (OECD).[19] In that document, the OECD promoted developments in artificial intelligence, while stressing the need to respect human rights and foster democratic values.[20]

At the conclusion of the PAL's workshop its President, Msgr. Vincenzo Paglia read Pope Francis's address to the PAL and participants.

---

[13] G20 Trade Ministers and Digital Economy Ministers, "G20 Ministerial Statement on Trade and Digital Economy," 2019, www.mofa.go.jp/files/000486596.pdf.

[14] This trio implies "responsible stewardship" aiming at "beneficial outcomes for people and the planet," i.e., "augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender, and other inequalities, and protecting natural environments" (G20 Trade Ministers and Digital Economy Ministers, "G20 Ministerial Statement," 11).

[15] These require respecting "freedom, dignity, and autonomy, privacy and data protection, nondiscrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights" as well as implementing "mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art" (G20 Trade Ministers and Digital Economy Ministers, "G20 Ministerial Statement," 11).

[16] These demand "transparency and responsible disclosure" regarding AI systems "to foster a general understanding of AI systems; to make stakeholders aware of their interactions with AI systems, including in the workplace; to enable those affected by an AI system to understand the outcome; and, to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation, or decision" (G20 Trade Ministers and Digital Economy Ministers, "G20 Ministerial Statement," 11).

[17] These urge "robust, secure, and safe" AI systems, avoiding any "unreasonable safety risk," ensuring "traceability, including in relation to datasets, processes, and decisions made" by these systems, and having in place a "systematic risk management approach … to address risks related to AI systems, including privacy, digital security, safety, and bias" (G20 Trade Ministers and Digital Economy Ministers, "G20 Ministerial Statement," 11–12).

[18] This stresses how "AI actors should be accountable for the proper functioning of AI systems" and for respecting these principles (G20 Trade Ministers and Digital Economy Ministers, "G20 Ministerial Statement," 12).

[19] See Organisation for Economic Cooperation and Development, *Recommendation of the Council on Artificial Intelligence*, OECD Legal Instruments 0449 (Paris: Organisation for Economic Cooperation and Development, 2020), 3; see also 7–8. The OECD is an intergovernmental economic organization with 36-member countries; it was founded in 1961 to stimulate economic progress and world trade. The US counts among its founding nations.

[20] For another example, see High-Level Expert Group on AI, *Ethics Guidelines for Trustworthy Artificial Intelligence* (Brussels: European Commission, 2019), 12–13. In this case, the principles are: respect for human autonomy, prevention of harm, fairness, and explicability.

In the Pope's text, we read that artificial intelligence "affects our way of understanding the world and ourselves. It is increasingly present in human activity and even in human decisions, and is thus altering the way we think and act" by informing human decisions.[21] Moreover, "on the socio-economic level, users are often reduced to 'consumers,' prey to private interests concentrated in the hands of a few. From digital traces scattered on the internet, algorithms now extract data that enable mental and relational habits to be controlled, for commercial or political ends, frequently without our knowledge."[22] Hence, for Francis, our freedom is challenged and "inequalities expand enormously; knowledge and wealth accumulate in a few hands with grave risks for democratic societies. Yet these dangers must not detract from the immense potential that new technologies offer. We find ourselves before a gift from God, a resource that can bear good fruits."[23]

For Pope Francis, the ethical agenda should be inclusive, involving "the human family as a whole"[24] and dialogical, leading to "identify paths of humanization, and thus of loving evangelization, that we can travel together. In this way we will be able to dialogue fruitfully with all those committed to human development, while keeping at the centre of knowledge and social praxis the human person in all his or her dimensions, including the spiritual."[25] While the Pope evokes the possibility of developing an "algor-ethics,"[26] he advocates for human rights, discernment, and the tenets of Catholic social teaching: the promotion of the common good, "the dignity of the person, justice, subsidiarity, and solidarity."[27] For Francis, these are the ethical resources that can accompany the current technological development of AI.

These themes shape Pope Francis's reflection on human agency, technology, and society. In his 2015 encyclical letter *Laudato Si'*, he appreciates the social benefits that technological developments made

---

[21] Francis, "Discorso ai Partecipanti alla Plenaria della Pontificia Accademia per la Vita Letto da S.E. Mons. Vincenzo Paglia, 28.02.2020," press.vatican.va/content/salastampa/it/bollettino/pubblico/2020/02/28/0134/00291.html#eng.

[22] Francis, " Discorso ai Partecipanti."

[23] Francis, " Discorso ai Partecipanti."

[24] Francis, " Discorso ai Partecipanti."

[25] Francis, " Discorso ai Partecipanti."

[26] "Algor-ethics" means "the ethical development of algorithms" (Francis, "Discorso ai Partecipanti"). See also Francis, "Address to Participants in the Congress on Child Dignity in the Digital World (November 14, 2019)," www.vatican.va/content/francesco/en/speeches/2019/november/documents/papa-francesco_20191114_convegno-child%20dignity.pdf.

[27] Francis, " Discorso ai Partecipanti." See also Antonio Spadaro and Paul Twomey, "Intelligenza Artificiale e Giustizia Sociale: Una Sfida per la Chiesa," *La Civiltà Cattolica* I, no. 4070 (2019): 121–31.

possible,[28] but is also aware of possible risks[29] and, in particular, of "the effects of technological innovations on employment, social exclusion, an inequitable distribution and consumption of energy and other services, social breakdown, increased violence, and a rise in new forms of social aggression, drug trafficking, growing drug use by young people, and the loss of identity" (no. 46).[30] Moreover, he worries about how human agency could be undermined by overemphasizing what he calls the technocratic paradigm that "tends to dominate economic and political life" and "exalts the concept of a subject who, using logical and rational procedures, progressively approaches and gains control over an external object" (nos. 109, 106).[31] According to Pope Francis, to reclaim agency, "We have to accept that technological products are not neutral, for they create a framework which ends up conditioning lifestyles and shaping social possibilities along the lines dictated by the interests of certain powerful groups. Decisions which may seem purely instrumental are in reality decisions about the kind of society we want to build" (no. 107).

Hence, the Pope calls for "an integral development and an improvement in the quality of life" (no. 46) that will "broaden our vision" (no. 112), address inequalities,[32] eliminate divisions,[33] and promote

---

[28] "Technology has remedied countless evils which used to harm and limit human beings" (*Laudato Si'*, no. 102). Moreover, "Technology is characteristic of the human being. It should not be understood as a force that is alien to and hostile to it, but as a product of its ingenuity through which it provides for the needs of living for oneself and for others. It is therefore a specifically human mode of inhabiting the world" (Francis, "Address to Participants in the Plenary Assembly of the Pontifical Academy for Life," 2019, www.vatican.va/content/francesco/en/speeches/2019/february/documents/papa-francesco_20190225_plenaria-accademia-vita.html). However, "There is an urgent need for greater study and discussion of the social effects of this technological development, for the sake of articulating an anthropological vision adequate to this epochal challenge" (Francis, "Address to Participants in the Plenary Assembly of the Pontifical Academy for Life," 2017, w2.vatican.va/content/francesco/en/speeches/2017/october/documents/papa-francesco_20171005_assemblea-pav.html).

[29] See Francis, "Address to Participants in the Congress on Child Dignity."

[30] See also Francis, "Message to the Executive Chairman of the 'World Economic Forum' on the Occasion of the Annual Gathering in Davos-Klosters (23–26 January 2018)," w2.vatican.va/content/francesco/en/messages/pont-messages/2018/documents/papa-francesco_20180112_messaggio-davos2018.html.

[31] For Pope Francis, "Our immense technological development has not been accompanied by a development in human responsibility, values, and conscience" (*Laudato Si'*, no. 105). On responsibility, see Francis, "Address to Participants in the Plenary Assembly," 2017, no. 2.

[32] See Francis, "Address to Participants in the Plenary Assembly," 2019.

[33] See Francis, "*Humana Communitas* (the Human Community): Letter of His Holiness Pope Francis to the President of the Pontifical Academy for Life for the 25th Anniversary of the Establishment of the Academy," 2019, www.vatican.va/content/francesco/en/letters/2019/documents/papa-francesco_20190106_lettera-accademia-vita.html.

freedom,[34] even to "limit and direct technology" by placing any development at the service of a type of progress "which is healthier, more human, more social, more integral" (no. 115).

Discernment[35] allows us to assess "the social effects of technological development"[36] and fosters a "general rethinking of social policies and human rights"[37] in order "to safeguard the dignity of the human person, in particular by offering to all people real opportunities for integral human development and by implementing economic policies that favour the family."[38]

Furthermore, "an ethic of sustainable and integral development, based on values that place the human person and his or her rights at the centre,"[39] rejects "a 'throwaway' culture and a mentality of indifference,"[40] and urges all people of good will to embrace and implement "a new vision aimed at promoting a humanism of fraternity and solidarity between individuals and peoples"[41] that includes caring for the whole planet, while being aware that "fraternity remains the unkept promise of modernity."[42]

Hence, "Artificial intelligence, robotics, and other technological innovations must be so employed that they contribute to the service of humanity and to the protection of our common home, rather than to the contrary, as some assessments unfortunately foresee."[43]

To sum up, Pope Francis invites us to consider technology by focusing on moral agents and agency, by considering which interests drive research and implementation of technological developments, and by empowering citizens with his inspired vision of integral development and a good society.

Agreeing on the importance of examining artificial intelligence in light of a moral vision that promotes agency, in what follows I discuss three ongoing implementations of AI within social contexts:[44] facial

---

[34] See *Laudato Si'*, no. 112. "Freedom and the protection of privacy are valuable goods that need to be balanced with the common good of society" (Francis, "Address to Participants in the Congress on Child Dignity").

[35] See Francis, "*Humana Communitas*," nos. 10–11. See also Francis, "Message to the Executive Chairman."

[36] Francis, "Address to Participants in the Plenary Assembly," 2017.

[37] Francis, "*Humana Communitas*."

[38] Francis, "Message to the Executive Chairman."

[39] Francis, "Message to the Executive Chairman."

[40] Francis, "Message to the Executive Chairman."

[41] Francis, "*Humana Communitas*," no. 6; see also no. 4; and Francis, "Address to Participants in the Congress on Child Dignity."

[42] Francis, "*Humana Communitas*," no. 13. See also Francis, "*Fratelli Tutti*: On Fraternity and Social Friendship," 2020, www.vatican.va/content/francesco/en/encyclicals/documents/papa-francesco_20201003_enciclica-fratelli-tutti.html.

[43] Francis, "Message to the Executive Chairman."

[44] For "an interdisciplinary research center dedicated to understanding the social implications of artificial intelligence," see New York University, "AI Now," 2020, ainowinstitute.org/. For initiatives in the Global South, see Abhishek Gupta and

recognition systems and how artificial intelligence is used, respectively, within the justice system and in workplaces. In relation to these specific contexts, the ethical agenda outlined by the *Rome Call for AI Ethics* and by Pope Francis could be further enriched. Hence, as I anticipated, an approach that critically examines these three implementations as forms of social control could first discern between an ethic of control and an ethic of risk, second revisit biopower and biopolitics, and third re-appropriate human-centered labor.

AI AND FACIAL RECOGNITION SYSTEMS: DIGITAL TRACKING

Within society, AI systems are increasingly present: from facial recognition services[45] to talking digital assistants—like Amazon Echo Plus (Alexa), Apple Homepod (Siri), and Google Home (Google Assistant);[46] from driverless cars undergoing driving testing;[47] to instant translation services like Google Translate.[48] These systems learn from enormous amounts of information. What are their ethical implications for individuals and society? I focus on facial recognition systems in law enforcement and security, as well as in public places and education.

Clearview AI: A Secretive Company[49]

Facial recognition systems in law enforcement are not new. Police departments have been using them for almost twenty years.[50] In the past, these systems searched only "government-provided images, such as mug shots and driver's license photos."[51] Now, they turn to the

---

Victoria Heath, "AI Ethics Groups Are Repeating One of Society's Classic Mistakes," *MIT Technology Review,* 2020, www.technologyreview.com/2020/09/14/1008323/ai-ethics-representation-artificial-intelligence-opinion/. I am grateful to Kristin E. Heyer for this last reference.

[45] See Cade Metz and Natasha Singer, "Newspaper Shooting Shows Widening Use of Facial Recognition by Authorities," *New York Times*, June 29, 2018, www.nytimes.com/2018/06/29/business/newspaper-shooting-facial-recognition.html.

[46] See Keith Collins and Cade Metz, "Alexa vs. Siri vs. Google: Which Can Carry on a Conversation Best?," *New York Times*, August 17, 2018, www.nytimes.com/interactive/2018/08/17/technology/alexa-siri-conversation.html.

[47] See Cade Metz, "Competing with the Giants in Race to Build Self-Driving Cars," *New York Times*, January 4, 2018, www.nytimes.com/2018/01/04/technology/self-driving-cars-aurora.html.

[48] See Gideon Lewis-Kraus, "The Great AI Awakening," *New York Times*, December 16, 2016, www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html.

[49] See Clearview AI, "Computer Vision for a Safer World," 2020, clearview.ai/.

[50] See Jennifer Valentino-DeVries, "How the Police Use Facial Recognition, and Where It Falls Short," *New York Times*, January 12, 2020, www.nytimes.com/2020/01/12/technology/facial-recognition-police.html.

[51] Kashmir Hill, "The Secretive Company That Might End Privacy as We Know It," *New York Times*, January 18, 2020, www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

facial recognition company Clearview AI, by taking a picture of a person and uploading it on the company's app. The app searches the Clearview database for public photos of that person, along with links to where those photos appear. The database has more than three billion images, scraped from Facebook, YouTube, Venmo, and millions of other websites—a practice that is ethically problematic, particularly when it concerns copyrighted data and personal information.[52] Every uploaded photo expands the Clearview database.

Clearview provides paid access to its app to hundreds of law enforcement agencies: from local police in Florida, the FBI and the Department of Homeland Security to Canadian law enforcement authorities. In 2019, "more than 600 law enforcement agencies have started using Clearview."[53] While federal and state law enforcement officers have "only limited knowledge of how Clearview works and who is behind it, they had used its app to help solve shoplifting, identity theft, credit card fraud, murder, and child sexual exploitation cases."[54] Clearview's business is larger than enforcement agencies, as it also includes "at least a handful of companies for security purposes."[55] Will the Clearview app—or other possible similar apps—be available to everyone who can pay, for whatever use they want to make of it?

Clearview claims that its app finds matches up to 75 percent of the time, but it is unclear how often there are false matches. The tool has not been tested by the National Institute of Standards and Technology—the federal agency that rates the performance of facial recognition algorithms.[56] In particular, "the larger the database, the larger the risk of misidentification because of the doppelgänger effect," which describes a non-biologically related look-alike of a living person.[57]

Without any public scrutiny, transparency, and accountability, "the tool could identify activists at a protest or an attractive stranger on the subway, revealing not just their names but where they lived, what they did and whom they knew."[58] Moreover, law enforcement agencies upload sensitive photos to servers of a "company whose ability to protect its data is untested."[59]

Clearview is using artificial intelligence to weaponize images available on the web, from social media to other websites. Curiously,

---

[52] See the European General Data Protection Regulation (GDPR), which came into force in May 2018, in "Complete Guide to GDPR Compliance," *GDPR.EU*, 2022, gdpr.eu/. See also "Web Scraping Laws," *TermsFeed*, 2021, www.termsfeed.com/blog/web-scraping-laws/.
[53] Hill, "The Secretive Company."
[54] Hill, "The Secretive Company."
[55] Hill, "The Secretive Company."
[56] See www.nist.gov/.
[57] Hill, "The Secretive Company."
[58] Hill, "The Secretive Company."
[59] Hill, "The Secretive Company."

the company depends on people's transparency and visibility, but it lacks transparency about its business practices and is almost invisible on the web.[60]

The Proliferation of Biased Facial Recognition Systems

While few US cities have banned government use of facial recognition (in California: San Francisco,[61] Oakland, and Berkeley; in Massachusetts: Brookline and Somerville), since 2018 some airports[62] and public venues, like Madison Square Garden in Manhattan,[63] have adopted it.

Lockport is a small city 20 miles east of Niagara Falls.[64] In the name of safety, in 2020 the Lockport School District installed a facial recognition system in its eight high schools "to help prevent mass shootings and stop sexual predators."[65] Hence, this is "the first known public school district in New York to adopt facial recognition, and one of the first in the nation."[66]

In higher education, Stanford University is already using facial recognition systems on its campus. Other universities might follow suit. However, at the University of Southern California, in Los Angeles, the planned implementation of facial recognition technology was cancelled due to backlash.[67]

---

[60] See Clearview AI, "Computer Vision for a Safer World."

[61] See Kate Conger, Richard Fausset, and Serge F. Kovaleski, "San Francisco Bans Facial Recognition Technology," *New York Times*, May 14, 2019, www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html.

[62] See Catie Edmondson, "An Airline Scans Your Face. You Take Off. But Few Rules Govern Where Your Data Goes," *New York Times*, August 6, 2018, www.nytimes.com/2018/08/06/us/politics/facial-recognition-airports-privacy.html.

[63] See Kevin Draper, "Madison Square Garden Has Used Face-Scanning Technology on Customers," *New York Times*, March 13, 2018, www.nytimes.com/2018/03/13/sports/facial-recognition-madison-square-garden.html.

[64] See Davey Alba, "Facial Recognition Moves into a New Front: Schools," *New York Times*, February 6, 2020, www.nytimes.com/2020/02/06/business/facial-recognition-schools.html.

[65] See Alba, "Facial Recognition," B6.

[66] Alba, "Facial Recognition," B1.

[67] See Sigal Samuel, "Is Your College Using Facial Recognition on You? Check This Scorecard," *Vox* 2020, www.vox.com/2020/1/29/21112212/facial-recognition-college-campus-scorecard; David Z. Morris, "College Backlash against Facial Recognition Technology Grows," *Fortune* 2020, fortune.com/2020/02/27/college-facial-recognition-technology-backlash/; Sameera Pant, Julia Shapero, and Saumya Gupta, "UCLA Decides Not to Implement Facial Recognition Technology after Student Backlash," *Daily Bruin*, 2020, dailybruin.com/2020/02/19/ucla-decides-not-to-implement-facial-recognition-technology-after-student-backlash; Drew Harwell, "Colleges Are Turning Students' Phones into Surveillance Machines, Tracking the Locations of Hundreds of Thousands," *The Washington Post*, December 24, 2019, www.washingtonpost.com/technology/2019/12/24/colleges-are-turning-students-

Globally, China is the leader in implementing facial recognition systems.[68] Within the country—in its cities and, in the future, even at crossroads in villages—cameras with facial recognition strictly control citizens, especially minorities like the Uyghurs—the Muslim Turkic-speaking minority in the Xinjiang Uyghur Autonomous Region in Northwest China.[69] China also leads in exporting and implementing these systems in the Global South:[70] from Singapore[71] to Mongolia; Ethiopia and Zimbabwe,[72] Kenya,[73] Uganda and Zambia;[74] Ecuador[75]

phones-into-surveillance-machines-tracking-locations-hundreds-thousands/. I am grateful to Peter Fay for these references.

[68] See Steven Feldstein, *The Global Expansion of AI Surveillance* (Washington, DC: Carnegie Endowment for International Peace, 2019). For a documentary, see Neil Docherty and David Fanning, "In the Age of AI," *PBS Frontline,* 2019, www.pbs.org/wgbh/frontline/film/in-the-age-of-ai/.

[69] See Charlie Campbell, "'The Entire System Is Designed to Suppress Us.' What the Chinese Surveillance State Means for the Rest of the World," *Time*, 2019, time.com/5735411/china-surveillance-privacy-issues/.

[70] See Mara Wang, "China's Dystopian Push to Revolutionize Surveillance," *The Washington Post*, August 18, 2017, www.washingtonpost.com/news/democracy-post/wp/2017/08/18/chinas-dystopian-push-to-revolutionize-surveillance/.

[71] See Alexa Hagerty and Igor Rubinov, "Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence," *arXiv* 2019, arxiv.org/abs/1907.07892.

[72] See Scott N. Romaniuk and Tobias Burgers, "How China's AI Technology Exports Are Seeding Surveillance Societies Globally," *The Diplomat,* 2018, thediplomat.com/2018/10/how-chinas-ai-technology-exports-are-seeding-surveillance-societies-globally/. See also Chinmayi Arun, "AI and the Global South: Designing for Other Worlds," in *The Oxford Handbook of Ethics of AI*, ed. M. D. Dubber, F. Pasquale, and S. Das (New York: Oxford University Press, 2020), 590–610, at 600.

[73] See Abdi Latif Dahir, "Chinese Firms Are Driving the Rise of AI Surveillance Across Africa," *Quartz Africa,* 2019, qz.com/africa/1711109/chinas-huawei-is-driving-ai-surveillance-tools-in-africa/. See also N. D. Francois, "Huawei's Surveillance Tech in Kenya: A Safe Bet?," *Africa Times,* 2019, africatimes.com/2019/12/18/huaweis-surveillance-tech-in-kenya-a-safe-bet/.

[74] See Joe Parkinson, Nicholas Bariyo, and Josh Chin, "Huawei Technicians Helped African Governments Spy on Political Opponents," *The Wall Street Journal*, August 15, 2019, www.wsj.com/articles/huawei-technicians-helped-african-governments-spy-on-political-opponents-11565793017.

[75] See Paul Mozur, Jonah M. Kessel, and Melissa Chan, "Made in China, Exported to the World: The Surveillance State," *New York Times*, April 24, 2020, www.nytimes.com/2019/04/24/technology/ecuador-surveillance-cameras-police-government.html. Similar surveillance systems have been sold to Venezuela, Bolivia, and Angola.

to Brazil[76] and Argentina.[77] AI technology makes possible social control, whether within China, as an expression of its authoritarian regime, or globally, by allowing Chinese access to these systems and their data, and by facilitating local authorities in their social control of citizens. Hence, reflections on AI should include systemic critiques of authoritarian states and of unethical policies harming democracies.[78]

Despite its increasing global implementation, facial recognition technology is not an exact science and it has always been controversial. The percentage of false matches is high.[79] While proponents view facial recognition as an important tool for catching criminals and tracking terrorists, critics are concerned about "privacy, accuracy, and racial bias."[80] In 2019, the National Institute of Standards and Technology tested 189 facial recognition algorithms from 99 developers.[81] The study found that algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces. The highest proportion of errors occurred in the case of Native Americans.[82]

---

[76] See Hagerty and Rubinov, "Global AI Ethics," 25. The authors refer to: Ray Walsh, "Brazil Employs Facial Recognition Technology to Tackle Crime Hotpots," *ProPrivacy,* 2019, proprivacy.com/privacy-news/brazil-facial-recognition-cameras; Chris Burt, "Possibility of Chinese Facial Biometrics Systems in Brazilian CCTV Network Raises Concerns," *Biometric Update,* 2019, www.biometricupdate.com/201901/possibility-of-chinese-facial-biometrics-systems-in-brazilian-cctv-network-raises-concerns.

[77] See Jose Hermosa, "Chinese Regime to Install Giant Surveillance System in Argentina," *The BL,* 2019, thebl.com/world-news/chinese-regime-to-install-giant-surveillance-system-in-argentina.html.

[78] See Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (New York: PublicAffairs, 2019).

[79] See Joy Buolamwini, Vicente Ordóñez, Jamie Morgenstern, and Erik Learned-Miller, *Facial Recognition Technologies: A Primer* (n.p.: Algorithmic Justice League, 2020); Natasha Singer, "Amazon Is Pushing Facial Technology That a Study Says Could Be Biased," *New York Times*, January 24, 2019, www.nytimes.com/2019/01/24/technology/amazon-facial-technology-study.html.

[80] Alba, "Facial Recognition," B6.

[81] The developers "included systems from Microsoft, biometric technology companies like Cognitec, and Megvii, an artificial intelligence company in China. The agency did not test systems from Amazon, Apple, Facebook, and Google because they did not submit their algorithms for the federal study" (Natasha Singer and Cade Metz, "Many Facial-Recognition Systems Are Biased, Says US Study," *New York Times*, December 19, 2019, www.nytimes.com/2019/12/19/technology/facial-recognition-bias.html).

[82] On bias in AI used in healthcare, see Tom Simonite, "A Health Care Algorithm Offered Less Care to Black Patients," *Wired,* 2019, www.wired.com/story/how-algorithm-favored-whites-over-blacks-health-care/. I am grateful to Mark McKenna for this reference. See also Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* 366, no. 6464 (2019): 447–53; Michele Samorani, Shannon Harris, Linda Goler Blount, Haibing Lu, and Michael A. Santoro, "Overbooked and Overlooked: Machine Learning and Racial Bias in Medical Appointment Scheduling," *SSRN* 2020, dx.doi.org/10.2139/ssrn.3467047; Nicole

The technology had more difficulty recognizing women than men—in particular African-American—and, in terms of age, "It falsely identified older adults up to 10 times more than middle-aged adults."[83] As Niraj Chokshi writes, "The problem, in part, is that facial recognition is only as good as the examples on which it is trained. And one widely used data set was estimated to be more than 75 percent male and more than 80 percent white."[84]

The technology's biases and lack of accuracy should be addressed and eliminated. Both a moratorium on the implementation of biometric technology in public spaces and appropriate ethical assessment and legal safeguards, are also needed.[85] Furthermore, neither the deep learning AI algorithms used for facial recognition systems, nor their applications are sufficiently critically evaluated. For civil liberties experts, "The technology—which can be used to track people at a distance without their knowledge—has the potential to lead to ubiquitous surveillance, chilling freedom of movement and speech."[86]

Using such a biased and error prone technology in civil society and law enforcement could lead to false accusations. In the US, people should be protected by a strong federal privacy law. Some citizens began to demand that facial recognition be regulated, to control those who control us.[87] Others already asked to ban it. Woodrow Hartzog,

---

Martinez-Martin, "What Are Important Ethical Implications of Using Facial Recognition Technology in Health Care?," *AMA Journal of Ethics* 21, no. 2 (2019): E180–87.

[83] Singer and Metz, "Many Facial-Recognition Systems." See also Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018): 77–91.

[84] Niraj Chokshi, "Facial Recognition's Many Controversies, from Stadium Surveillance to Racist Software," *New York Times*, May 15, 2019, www.nytimes.com/2019/05/15/business/facial-recognition-software-controversy.html.

[85] See Davide Castelvecchi, "Beating Biometric Bias," *Nature* 587, no. 7834 (2020): 347–49, at 348; Kate Crawford, "Regulate Facial-Recognition Technology," *Nature* 572, no. 7771 (2019): 565; Richard Van Noorden, "The Ethical Questions That Haunt Facial-Recognition Research," *Nature* 587, no. 7834 (2020): 354–58.

[86] Singer and Metz, "Many Facial-Recognition Systems." "AI-driven technologies have a pattern of entrenching social divides and exacerbating social inequality, particularly among historically-marginalized groups" (Hagerty and Rubinov, "Global AI Ethics," 1). For a movement towards equitable and accountable AI, see Algorithmic Justice League (AJL): www.ajl.org/. AJL was founded by computer scientist Joy Buolamwini at the Massachusetts Institute of Technology in Cambridge, MA.

[87] Because of citizens' pressure, animated by Alistair McTaggart's engagement, California approved the Consumer Privacy Act (2018) granting consumers more control over their personal information collected by businesses (California Legislative, "California Consumer Privacy Act, Title 1.81.5," 2018, leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5). The measure was implemented on January 1, 2020. In Europe, General Data Protection Regulation is in place, but it does not clarify whether

professor of law and computer science at Northeastern University, tells poignantly: "I don't see a future where we harness the benefits of face recognition technology without the crippling abuse of the surveillance that comes with it. The only way to stop it is to ban it."[88] One might wonder whether this will ever happen. The powerful, largely hidden effects of algorithms in American life enhance biases and discriminations that already characterize our social fabric with its racial and gender inequities.

AI AND THE JUSTICE SYSTEM: AUTOMATED JUSTICE

In the US, at the federal and state levels, as well as in at least sixteen European countries, the justice system too is increasingly relying on artificial intelligence.[89] In almost every US state,[90] the most commonly used algorithms—called "pretrial risk assessment" or "risk assessments" or "evidence-based methods"—claim to predict future behavior of defendants and incarcerated persons. These AI systems are supposed to estimate the likelihood that the defendant will re-offend before trial (recidivism risk) and the likelihood the defendant will fail to appear at trial (FTA).[91]

Moreover, these algorithms are used to set bail, determine sentences, and even assess one's guilt or innocence. Yet we do not know how these systems work. For Aleš Završnik, "The technical sophistication of the new AI systems used in decision-making processes in criminal justice settings often leads to a 'black box' effect. The intermediate phases in the process of reaching a decision are by definition hidden from human oversight due to the technical complexity involved."[92] Hence, transparency, comprehensibility, and explainability are lacking in crucial decision-making processes.

To make a "criminal risk assessment," the algorithms consider personal characteristics like age, sex, geography, family background, and

researchers can collect photos of people for their research without their consent. See gdpr-info.eu/.

[88] Hill, "The Secretive Company."

[89] See Cade Metz and Adam Satariano, "An Algorithm That Grants Freedom, or Takes It Away," *New York Times*, February 6, 2020, www.nytimes.com/2020/02/06/technology/predictive-algorithms-crime.html.

[90] "In Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington, and Wisconsin, the results of such assessments are given to judges during criminal sentencing" (Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," *ProPublica,* May 23, 2016, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing).

[91] See Epic, "Algorithms in the Criminal Justice System: Risk Assessment Tools," *epic.org,* 2020, epic.org/algorithmic-transparency/crim-justice/.

[92] Aleš Završnik, "Criminal Justice, Artificial Intelligence Systems, and Human Rights," *ERA Forum* 20 (2020): 568.

employment status.[93] Hence, "As a result, two people accused of the same crime may receive sharply different bail or sentencing outcomes based on inputs that are beyond their control—but have no way of assessing or challenging the results."[94]

The algorithms are trained by relying on historical crime data. In such a way, the AI system supposedly identifies crime patterns. Karen Hao, however, stresses how

> those patterns are statistical *correlations—nowhere near the same as causations*. If an algorithm found, for example, that low income was correlated with high recidivism, it would leave you none the wiser about whether low income actually caused crime. But this is precisely what risk assessment tools do: they turn correlative insights into causal scoring mechanisms. Now populations that have historically been disproportionately targeted by law enforcement—especially low-income and minority communities—are at risk of being slapped with high recidivism scores. As a result, the algorithm could amplify and perpetuate embedded biases and generate even more bias-tainted data to feed a vicious cycle.[95]

These algorithms render the justice system less fair for criminal defendants because these technologies "are largely privately owned and sold for profit. The developers tend to view their technologies as trade secrets. As a result, they often refuse to disclose details about how their tools work, including to criminal defendants and their attorneys, even under a protective order, even in the controlled context of a criminal proceeding or parole hearing."[96]

Despite these limitations, predictive algorithms are spreading. In the US, authorities use them to set police patrols, prison sentences, and probation rules; in the Netherlands, to flag welfare fraud risks; and, in the UK, to rate which teenagers could become criminals. At the same time, "United Nations investigators, civil rights lawyers, labor unions and community organizers have been pushing back."[97] Algorithms could contribute to grant our freedom or take it away.[98]

---

[93] See Derek Thompson, "Should We Be Afraid of AI in the Criminal-Justice System?," *The Atlantic,* 2019, www.theatlantic.com/ideas/archive/2019/06/should-we-be-afraid-of-ai-in-the-criminal-justice-system/592084/.
[94] See Epic, "Algorithms in the Criminal Justice System."
[95] Karen Hao, "AI Is Sending People to Jail—and Getting It Wrong," *MIT Technology Review,* 2019, www.technologyreview.com/2019/01/21/ 137783/algorithms-criminal-justice-ai/.
[96] Rebecca Wexler, "When a Computer Program Keeps You in Jail," *New York Times*, June 13, 2017, www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html.
[97] Metz and Satariano, "An Algorithm That Grants Freedom."
[98] See Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (New York: Basic Books, 2015).

AI AND THE JOB MARKET: CHANGING INVESTMENTS, PRODUCTION, AND MARKETING

With a very emphatic and optimistic tone, the multinational Accenture claims that "in 12 developed economies AI could double annual economic growth rates in 2035 by changing the nature of work and creating a new relationship between man [*sic*] and machine. The impact of AI technologies on business is projected to increase labor productivity by up to 40 percent and enable people to make more efficient use of their time."[99] However, no indication of the social costs and transformations that AI will require, nor any comment on what will happen to the other 183 economies is provided.

In its 2017 report *Jobs Lost, Jobs Gained*, the McKinsey Global Institute provides a more nuanced assessment by stressing that by 2030, up to one third of the American workforce will need to change occupation.[100] Technological progress will lead these changes, but economic policies and social attitudes are no less relevant. History shows how humankind adapted to major technological changes (e.g., electricity and computers). With AI, transformations in workplaces might occur at a faster pace than in the past, causing greater disruption.

AI is rapidly introducing multiple levels of automation in the workplace. In the case of the hiring process, "Designed by the recruiting-technology firm HireVue, the [AI] system uses candidates' computer or cellphone cameras to analyze their facial movements, word choice, and speaking voice before ranking them against other applicants based on an automatically generated 'employability' score.… More than 100 employers now use the system, including Hilton and Unilever, and more than a million job seekers have been analyzed."[101] How facial expressions and emotions are assessed and evaluated,[102] and which criteria the AI system uses to select job candidates remains unclear. As a response, "In August [2019], Illinois Gov. J. B. Pritzker (D) signed a first-in-the-nation law that will force employers to tell job

---

[99] Accenture, "Artificial Intelligence," 2020, www.accenture.com/us-en/insights/artificial-intelligence-summary-index.

[100] See James Manyika, Susan Lund, Michael Chui, Jacques Bughin, Jonathan Woetzel, Parul Batra, Ryan Ko, and Saurabh Sanghvi, *Jobs Lost, Jobs Gained: What the Future of Work Will Mean for Jobs, Skills, and Wages* (San Francisco: McKinsey Global Institute, 2017).

[101] Drew Harwell, "A Face-Scanning Algorithm Increasingly Decides Whether You Deserve the Job," *The Washington Post*, October 22, 2019, www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/.

[102] On misinterpreting emotions and facial expressions, see Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak, "Emotional Expressions Reconsidered: Challenges to Inferring Emotion from Human Facial Movements," *Psychological Science in the Public Interest* 20, no. 1 (2019): 1–68.

applicants how their AI-hiring system works and get their consent before running them through the test."[103]

In warehouses, artificial intelligence already changed the working conditions. Online retailers like Amazon rely on AI for each step of the retail experience. Until recently, robots were able only to perform simple, repetitive motions, like picking up boxes. Boxes are easy objects because they have set dimensions that do not change. New robots, however, are more selective. They can pick up very different objects. For example, inside a "German warehouse, the robot can pick and sort more than 10,000 different items, and it does this with more than 99 percent accuracy."[104]

These changes in automation will influence employment and the whole marketplace. Both blue-collar and white-collar jobs will be lost, beginning with more socially vulnerable workers—among them, the elderly and women. For example, "The new warehouses will be built around A.I. robots and not humans."[105] To compensate and balance the loss of these jobs, new jobs are created. In the US, in the last six years, retail job losses "have been almost exactly counterbalanced by a gain of 118,000 light-truck or delivery-service driver jobs. The number of heavy-truck and tractor-trailer drivers increased more than 175,000 over the same period, making these two driving jobs among the fastest-growing occupations in the United States."[106]

Citizens need to reflect, however, on whether these new jobs offer the same protections and benefits that workers were able to negotiate in other productive sectors and jobs. For example, "Amazon uses AI to calculate how many drivers are needed in an area at any moment, based on package volume, weight, and travel time."[107] Working hours are flexible, with more workers hired during the holiday seasons, but "Drivers are responsible for providing their own vehicle, as well as fuel and other expenses. There are no benefits, little job security, and reports of sometimes grueling working conditions."[108]

A further example concerns India, plagued by the caste system. As Isabel Wilkerson has powerfully argued, the caste system is tragically responsible for the suffering and the inhuman social and working

---

[103] Harwell, "A Face-Scanning Algorithm." The measure was implemented on January 1, 2020.
[104] Adam Satariano and Cade Metz, "A Warehouse Robot Learns to Sort out the Tricky Stuff," *New York Times*, January 29, 2020, www.nytimes.com/2020/01/29/technology/warehouse-robot.html.
[105] Satariano and Metz, "A Warehouse Robot."
[106] David Deming, "The Robots Are Coming: Prepare for Trouble," *New York Times*, January 30, 2020, www.nytimes.com/2020/01/30/business/artificial-intelligence-robots-retail.html.
[107] Deming, "The Robots Are Coming."
[108] Deming, "The Robots Are Coming."

conditions of too many people in India and elsewhere.[109] In particular, the oppression, marginalization, discrimination, and stigma that the Dalits experience is inhuman. What they suffer is unacceptable and should not be tolerated. Transformative social justice is urgent. In 2014, struggling to improve the country's sanitation, the Prime Minister Narendra Modi started "Swachh Bharat Abhiyan" (Clean India Mission), a campaign aimed at eliminating open defecation and improve human waste management. No mention was made of the urgent need to abolish manual scavenging of human waste performed daily by many Dalit women and men—a job no human being should ever do.[110]

The recent launch of Bandicott—an AI-controlled robot that replaces manual scavenging—is promoted by social activists attempting to abolish the caste-based labor of manual scavenging. Paradoxically, if these robots are implemented, the Dalits will lose those jobs and their social exclusion will further increase the lack of provisions to secure humanly appropriate jobs.[111] Hence, implementing Bandicott is not sufficient. Human dignity requires a more committed engagement to eliminate the caste system. The dignity of work demands to change unjust social structures and inhuman working conditions. The Dalits should contribute to the country's social development with humane jobs that do not threaten their health and well-being and are fairly compensated. They are citizens with equal rights deserving social recognition and respect.

To address the terrible working conditions that characterized the Industrial Revolution, "Beginning in the early 20th century, trade unions and new government regulations acted together to raise pay, improve working conditions, and increase workplace safety."[112] Hence, we need to prevent exploitation by protecting workers' rights and people's working conditions. According to David Deming, director of the Malcolm Wiener Center for Social Policy at the Harvard Kennedy

---

[109] See Isabel Wilkerson, *Caste: The Origins of Our Discontents* (New York: Random House, 2020). On caste in India and the USA, see her chapter 7: "Through the Fog of Delhi to the Parallels in India and America," 71–77.

[110] See Assa Doron and Robin Jeffrey, *Waste of a Nation: Garbage and Growth in India* (Cambridge, MA: Harvard University Press, 2018), 69–97. I am grateful to Dhinakaran Savariyar for this suggestion and reference.

[111] See Garima Bora, "A Robot to End Manual Scavenging? This Startup Can Provide the 'Swachh Bharat' We Need," *The Economic Times,* 2019, economictimes.indiatimes.com/small-biz/startups/features/a-robot-to-end-manual-scavenging-this-startup-can-provide-the-swachh-bharat-we-need/articleshow/69685536.cms; Malavika Prasad and Vidushi Marda, "Interrogating 'Smartness': A Case Study on the Caste and Gender Blind Spots of the Smart Sanitation Project in Pune, India," in *Artificial Intelligence: Human Rights, Social Justice, and Development*, ed. Global Information Society Watch (New York: Association for Progressive Communications, 2019), 145–51.

[112] Deming, "The Robots Are Coming."

School, "We need to accept that we cannot stop the coming wave of technological change. But we can moderate its impact on society. We should act with purpose, embracing AI as a tool that will enable us to create a better and fairer world."[113]

Are we getting ready to address these changes in the job market? Are we reflecting critically on how to rethink the human role in workplaces? Are we considering human labor as an irreplaceable personal and social dimension that characterizes individual and collective flourishing, integral to promoting the common good within society? How should we think about education and getting ready for the workplace? With others, Juliet Schor invites us to consider how and for what we work, what our future ways of working will be, and how we will balance work and free time.[114]

THEOLOGICAL DISCOURSE: ADDRESSING SOCIAL CONTROL

The areas of social presence of artificial intelligence briefly examined—from civil society to education and law enforcement, from the judicial system to the workplace—show how social control occurs in multiple and diversified ways. In these contexts, AI technology is used to implement controlling power dynamics that affect citizens and limit people's freedom and agency. Theological ethics can contribute to identifying forms of social control that inhibit personal and social flourishing. After briefly summing up specific ethical challenges, with focused ethical approaches theological discourse further enriches the ethical agenda by engaging each of these three diverse contexts in which AI is implemented.

First, the proliferating and expanding use of facial recognition systems—from law enforcement to education, and civil society—is ethically problematic.[115] Citizens are neither informed nor protected. A critical assessment of these forms of social control can rely on articulating ethics of control and of risk.

Second, profiling, biases, and stigmatization depending on race, ethnicity, geography, residence, history, age, economic conditions,

---

[113] Deming, "The Robots Are Coming."

[114] See Juliet B. Schor, *Plenitude: The New Economics of True Wealth* (New York: Penguin, 2010); Juliet Schor and Craig J. Thompson, eds., *Sustainable Lifestyles and the Quest for Plenitude: Case Studies of the New Economy* (New Haven, CT: Yale University Press, 2014); Juliet B. Schor, *After the Gig: How the Sharing Economy Got Hijacked and How to Win It Back* (Oakland, CA: University of California Press, 2020). To address changes in workplaces, some Northern European countries with strong welfare systems are considering reducing the number of working hours per week. They are debating whether each citizen should be paid a sufficient wage, even when jobs are not available, to protect their ability to live, buy, and consume.

[115] See Zaheer Allam and David S. Jones, "On the Coronavirus (COVID-19) Outbreak and the Smart City Network: Universal Data Sharing Standards Coupled with Artificial Intelligence (AI) to Benefit Urban Health Monitoring and Management," *Healthcare* 8, no. 1 (2020), 10.3390/healthcare8010046.

political and religious beliefs, bodily shape, height, weight, and class seem to inform the uses of artificial intelligence in the justice system. Paradoxically, AI could affect us by controlling us, while its stated purpose is to avoid any abuse that could harm us (from crimes to school shootings and terrorist attacks within the nation and internationally). Fear seems to dominate how AI is used in these social contexts. Because we are afraid of what could happen, as a civil society we might let our fear decide and take away hard won liberties and rights. Critical reflections that unmask and redirect biopower and biopolitics seem to be appropriate.

Third, AI is already changing our workplaces and leading a new technological revolution. The automation that AI is introducing requires a different expertise. Old jobs will be lost and new types of jobs will be created. To be 21st century Luddites and fight strenuously against technological transformations is neither intelligent nor wise. Imagination and creativity are needed to train current and future workers by protecting workers' rights and benefits. Hence, contributions that inform our reflection on human labor, working arrangements, and workplaces are beneficial.

Discerning between an Ethic of Control and an Ethic of Risk

Theological discourse should, first, identify any biased attempt and logic aimed at realizing oppressive social control in ways that disempower moral agents, as well as their social presence and action, by acquiring, storing, and manipulating any information that concerns them. Informing people that facial recognition systems are in place and gather data, and asking for one's consent to collect, store, and use data are essential. However, provision of information and request for consent are far from being implemented.

Per se, limited and regulated forms of social control are not necessarily evil practices. Any type of social control should be justified, respect citizens' privacy, protect their liberties, and avoid any racial disparity, bias, and discrimination. The rule of law, law enforcement, and public health measures—to contain the spreading of infections and protect the health of citizens—exemplify three contexts in which defined, bound, and limited social control aims at serving the citizens' quality of life.[116]

To express this concept differently, power is not necessarily abusive, but too often it is abused. To play with words, when power leads to oppressive social control, which discriminates unjustly among citizens, it should be controlled. For Sharon Welch, however, control

[116] Tragically, however, as social events continue to remind us, the rule of law, law enforcement, and public health measures continue to be abused and serve forms of social control that harm vulnerable citizens. I am grateful to Aimee Hein for stressing this important point.

might pervert moral agency because "our moral and political imagination is shaped by an ethic of control, a construction of agency, responsibility, and goodness which assumes that it is possible to guarantee the efficacy of one's actions."[117]

In anthropological terms, decisions and actions that embody and foster oppressive social control seem to be informed, again, by a pervasive and paralyzing fear that influences individual and social actions.[118] What is feared is perceived as a threatening "other," whether in the case of moral agents—i.e., isolated human beings, groups, and institutions—or technological advances and the opportunities they might offer. A further dimension of such a fear is the inability to appreciate how moral agents are capable of performing responsible actions. Hence, what is feared is people's ability to use technology in virtuous ways, critically examined and aimed at promoting individual and social flourishing—what could be defined as virtuous social control.

Second, by stressing the relational dimensions that constitute the social fabric, theological discourse should foster virtuous social dynamics regulating the use of technologies by placing them at the service of a social life that empowers citizens and promotes their social well-being. Sharon Welch, Cynthia Crysdale, and Kristin Heyer exemplify authors who help us to aim at this goal by discerning between an ethic of control and an ethic of risk.

Because human beings are relational beings, created in the image of God,[119] an ethic of risk starts with the risk associated with one's being and with engaging in relationships. The unpredictability of everyday life, with its multiple and multifaceted relationships, is assumed and lived without the intent of controlling each of its dimensions and aspects. We recognize what can generate fear and is ethically risky. The ethical response avoids embracing an attitude of controlling domination, flawed because it lures us with the unrealistic goal of

---

[117] Sharon D. Welch, *A Feminist Ethic of Risk*, rev. ed. (Minneapolis, MN: Fortress, 2000), 14. Quoted in Andrea Vicini, SJ, "Ethical Issues and Approaches in Stem Cell Research: From International Insights to a Proposal," *Journal of the Society of Christian Ethics* 23, no. 1 (2003): 98, n. 113.

[118] For a biblical study, see Bruna Costacurta, *La Vita Minacciata: Il Tema della Paura nella Bibbia Ebraica*, Analecta Biblica (Roma: Pontificio Istituto Biblico, 1988). For a pastoral approach, see Virginio Spicacci, *Gesù, Una Buona Notizia! Vols. 1-2*, Formazione (Roma: Apostolato della Preghera, 2015).

[119] See Mary Jo Iozzio, "Radical Dependence and the *Imago Dei*: Bioethical Implications of Access to Healthcare for People with Disabilities," *Christian Bioethics* 23, no. 3 (2017): 234–60; Chammah Judex Kaunda, "Bemba Mystico-Relationality and the Possibility of Artificial General Intelligence (AGI) Participation in *Imago Dei*," *Zygon* 55, no. 2 (2020): 327–43; Karen O'Donnell, "Performing the *Imago Dei*: Human Enhancement, Artificial Intelligence and Optative Image-Bearing," *International Journal for the Study of the Christian Church* 18, no. 1 (2018): 4–15.

eliminating any uncontrolled element and factor, as well as any perceived danger and risk, and seduces with the fake panacea of total control.[120]

An ethic of risk addresses scientific issues that concern individuals and society by relying on an ongoing discerning attitude evaluating whether to pursue research and implement its applications in society. Such an ethic formulates ethical questions, suggests caution when necessary, and examines possible alternatives when choices are due. An ethic of risk is not risky; it invites us to identify virtuous ways and engage virtuously in what might be perceived as somehow ethically risky. Prudent discernment is at the core of an ethic of risk.

Other elements characterize an ethic of risk. For Sharon Welch, a feminist ethic of risk implies "a redefinition of responsible action, grounding in community, and strategic risk-taking."[121] In particular, responsible action means "the creation of a matrix in which further actions are possible, the creation of the conditions of possibility for desired ends."[122] Stressing its communal dimension, an ethic of risk aims at promoting relational and institutional dynamics within the social context without the intent of fostering manipulative social control and relying on the ethical empowerment of all moral agents. Strategic risk-taking implies that an ethic of risk exposes the vulnerability of moral agents by presupposing continuing discernment and evaluation without offering the apparent warranties of an ethic of control. The ethic of risk presupposes the human ability of addressing what appears to be a risk—for individuals and society—in ways that do not foster unnecessary risk-taking and neither harm moral agents ethically, emotionally, psychologically, spiritually, and socially, nor inhibit their personal and social agency.[123]

While Welch advocates for an ethic of risk as radical response to the limits and faults of an ethic of control, both Cynthia Crysdale and Kristin Heyer stress the helpful tension between an ethic of control and an ethic of risk. For Crysdale, "The goal of one's actions may be to enhance control for those who lack it, but this goal will be undertaken in a stance of risk," that is, marked by "redefinition of responsible action, grounding in community, and strategic resourcefulness over the long haul."[124]

---

[120] On totality as ethically problematic, see Emmanuel Levinas, *Totalité et infini: essai sur l'extériorité*, 3rd ed., Phaenomenologica (La Haye: Nijhoff, 1968).

[121] Welch, *A Feminist Ethic of Risk*, 46. Quoted in Vicini, "Ethical Issues and Approaches," 84.

[122] Welch, *A Feminist Ethic of Risk*, 46.

[123] For an ethic of risk on war and peace, see Sharon D. Welch, *After Empire: The Art and Ethos of Enduring Peace* (Minneapolis, MN: Fortress, 2004), 159–84.

[124] Cynthia S. W. Crysdale, "Making a Way by Walking: Risk, Control, and Emergent Probability," *Theoforum* 39, no. 1 (2008): 57. Quoted in Kristin E. Heyer, "The Social Witness and Theo-Political Imagination of the Movements: Creating a New Social

Being attentive to existing conditions of social inequities and how they inhibit moral agency in the social fabric, for Heyer an ethic of risk "acknowledges that actions may only lead to partial results, but amid a long-term struggle with oppressive situations, the goal of moral action is 'the creation of new conditions of possibility for the future.'"[125] On the one hand, "An ethic of risk thus entails redefining responsible action, in terms of 'the creation of the conditions of possibility for desired changes.'"[126] On the other hand, moral agency is understood "in terms of 'responsible actions within the limits of bounded power,' entailing 'persistent defiance and resistance in the face of repeated defeats.'"[127] For Heyer, "Integrating a view of moral agency as entailing both control and risk seeks to engender contexts and conditions that empower agents, while attending to the responsibilities of the vulnerable and those whose choices impact them."[128]

While forms of social control might perceive moral agents as threats that should be manipulated and disempowered, an ethic of control aims at empowering citizens. Hence, sinful dynamics and practices that use AI and in particular facial recognition systems for ethically problematic purposes, motivated by fear and for the sake of social control, should receive citizens' attention. Echoing what Crysdale and Heyer suggest, empowered citizens should implement forms of democratically supervised social control. By such means, moral agents would express the tension between an ethic of control and of risk in ways that identify, name, and regulate uses of facial recognition systems that do not harm citizens, with great attention to those who in society are more vulnerable. Virtuous social agency is possible and virtuous social practices are needed.

Revisiting Biopower and Biopolitics

Initially proposed by the French philosopher Michel Foucault (1926–1984), the notion of "biopower" leads to critically examine technologies that affect personal and social life by focusing on the dynamics of power and their effects concentrated on people's bodies.[129]

Space as a Challenge to Catholic Social Thought," *Journal of Catholic Social Thought* 10, no. 2 (2013): 326.

[125] Heyer, "The Social Witness," 325. She quotes Crysdale, "Making a Way by Walking," 40–41. See also Kristin E. Heyer, *Kinship across Borders: A Christian Ethic of Immigration* (Washington, DC: Georgetown University Press, 2012), 155.

[126] Heyer, "The Social Witness," 325, quoting Welch, *A Feminist Ethic of Risk*, 46. See also Heyer, *Kinship across Borders*, 155.

[127] Heyer, "The Social Witness," 325.

[128] Heyer, "The Social Witness," 326.

[129] See Michel Foucault, *The Birth of the Clinic: An Archaeology of Medical Perception*, trans. A. M. S. Smith, World of Man (New York: Pantheon, 1973). See also Black Hawk Hancock, "Michel Foucault and the Problematics of Power: Theorizing DTCA and Medicalized Subjectivity," *Journal of Medicine and*

Inspired by French writer Georges Bataille (1897–1962), Foucault's "biopolitics" traces political arrangements and practices through which biopower is exercised over populations and people, acting on their bodies.[130] In social contexts, biopower and biopolitics allow us to examine the specific techniques that, in their multiple forms and contexts, are implemented. These techniques concern human bodies and people's life stories throughout their life span from birth to death. Without being simply critical tools, biopower and biopolitics also encompass the need to identify, unmask, unveil, and name ethically problematic social dynamics. At the same time, biopower and biopolitics should empower resistance and transformative processes, even in the case of AI used in the justice system.[131]

In particular, as a form of social control, biopower seeks to define what is considered as "normal" and socially acceptable by those who are in positions of power. This process of "normalization" neither depends on predetermined "rationales" informed by principles nor manifests a virtuous moral life (e.g., by identifying what it means to be a virtuous human being), nor has any intention to protect and promote the dignity of people and populations.[132] On the contrary, normalizing

---

Philosophy 43, no. 4 (2018): 439–68. For an overview, see Andrea Vicini, SJ, "Biopotere," *Aggiornamenti Sociali* 61, no. 1 (2010): 61–64.

[130] See Michel Foucault, *The Birth of Biopolitics: Lectures at the College de France, 1978–1979* (New York: Graham Burcell, 2008). See also Jeferson Bertolini, "O Conceito de Biopolítica em Foucault: Apontamentos Bibliográficos," *Revista Missioneira* 21, no. 1 (2019): 75–91; Isacco Turina, "Vatican Biopolitics," *Social Compass* 60, no. 1 (2013): 134–51; Stephen R. Schloesser, SJ, "*Dancing on the Edge of the Volcano*: Biopolitics and What Happened after Vatican II," in *From Vatican II to Pope Francis: Charting a Catholic Future*, ed. P. Crowley, SJ (Maryknoll, NY: Orbis, 2014), 3–26.

[131] See Antoaneta Roussi, "Resisting the Rise of Facial Recognition," *Nature* 587, no. 7834 (2020): 350–53.

[132] For further developments, see contributions in disability studies. For example, Lennard J. Davis, *The End of Normal: Identity in a Biocultural Era* (Ann Arbor, MI: University of Michigan Press, 2014); Lennard J. Davis, "Introduction: Disability, Normality, and Power," in *The Disability Studies Reader*, ed. L. J. Davis, 5th ed. (New York: Routledge, 2017), 1–15; Lennard J. Davis, *Enforcing Normalcy: Disability, Deafness, and the Body* (London: Verso, 1995); Richard Cross, "Impairment, Normalcy, and a Social Theory of Disability," *Res Philosophica* 93, no. 4 (2016): 693–714; Jos V. M. Welie, "Persons with Intellectual and Developmental Disabilities: Philosophical Reflections on Normalcy, Disability, and the *Imago Dei*," *Journal of Religion & Society*, Supplement Series, 12 (2015): 13–38; Rosemarie Garland-Thomson, "Disability Bioethics: From Theory to Practice," *Kennedy Institute of Ethics Journal* 27, no. 2 (2017): 323–39; Rosemarie Garland-Thomson, "The Case for Conserving Disability," *Journal of Bioethical Inquiry* 9, no. 3 (2012): 339–55; Chandra Kavanagh, "What Contemporary Models of Disability Miss: The Case for a Phenomenological Hermeneutic Analysis," *International Journal of Feminist Approaches to Bioethics* 11, no. 2 (2018): 63–82. See also the ten contributions in the special issue "Engaging Disability," edited by M. J. Romero and M. J. Iozzio, of the *Journal of Moral Theology* 6, no. 2 (2017).

biopower aims at satisfying the drive to control, transforming subjects into objects.

Biopower, and biopolitics as its correlative political expression, foster manipulative attitudes that neither pay attention to moral subjects nor have any consideration or respect for values, with the human potential and the capabilities they express, for cultural, religious, and historical contexts with their specificity,[133] or the social networks to which people belong.

Within the justice system, the uses of AI discussed manifest how new technological tools can support longstanding discriminatory attitudes by expressing biopower in new ways, focusing on individuals, and their social presence and action—in their past, present, and future—to control their bodies and influence their agency.

More recently, other authors have further developed the understanding of biopower. For example, the Italian philosopher Giorgio Agamben applies the notion of biopower to the entire sphere of sovereignty, noting that sovereign power is imposed not only on subjects as holders of rights, but on the "naked life" of people—understood as what is exposed to the violence of that power.[134] The tragic, emblematic example is the Nazi racist dictatorship, which used medicalized power to exercise total control over the body of their victims.

Roberto Esposito, political and moral philosopher, interprets the biopower present in biopolitics by using the category of *bíos*: a form of political life—i.e., a community (*communitas*)—that emerges from the dynamics of "immunization," but that is not determined by them.[135] The willingness of becoming immune to the "other" is the basis of biopower and biopolitics and is evident in how people defend themselves against everything that is "other," because the "other" is perceived as potential threat. To respond and resist, Esposito proposes to avoid what fear would suggest—i.e., total closure to the "other" who is considered an outsider—and to strengthen effectively one's community. For Esposito, *communitas* is an example of a social context shaped by positive dynamics and relationships. *Communitas* manifests the positive results that can be experienced in the encounter between political dynamics and human life, leading to personal and social flourishing—what he defines as *bíos*.

---

[133] On AI in diverse cultural and religious contexts, see Antonio Spadaro, SJ, and Thomas Banchoff, "Intelligenza Artificiale e Persona Umana: Prospettive Cinesi e Occidentali," *La Civiltà Cattolica* II, no. 4055 (2019): 432–43.

[134] Giorgio Agamben, *Il Potere Sovrano e la Nuda Vita*, Homo Sacer (Torino: Giulio Einaudi, 1995). See also Georges De Schrijver, SJ, "Giorgio Agamben's Analysis of the Mechanism of Exclusion or the Logic of Sovereign Power," *Budhi: A Journal of Ideas and Culture* 18, no. 3 (2014): 1–18.

[135] See Roberto Esposito, *Bíos: Biopolitics and Philosophy*, trans. T. Campbell, Posthumanities Series (Minneapolis, MN: University of Minnesota Press, 2008).

Finally, rather than describing biopower as an inclusive and all-encompassing notion, American anthropologist Paul Rabinow and English sociologist Nikolas Rose focus on three dynamics. For them, in matters of life and death biopower occurs first when people and authorities state "their truth," second when they foster practices aimed at controlling others, and third when they promote dependence.[136] Biopower is manifested in oppressive discourses and abusive practices.

Social control on our bodies happens also beyond facial recognition and outside the justice system. In India, for example, Aadhaar is "the biometrics-based 'unique identity' number database" designed by the "software billionaire, Nandan Nilekani" as mandatory "for anyone who wants to access the Indian welfare system."[137] Due to its malfunctions and because "enrolling in the database will not spare an impoverished person the effort of opening a bank account, or acquiring a ration card … Aadhaar has played havoc with people's lives and has caused people to starve by preventing them from accessing the government services that deliver their basic right to food."[138] Finally, "The architecture of the biometric data collection system does not account for what happens to their bodies as a result of living on the streets."[139]

The variations in emphases among authors interpreting biopower and biopolitics, as well as this Indian example, suggest the need for urgent and careful discernment. Biopower and biopolitics allow us to examine critically how AI controls citizens and influences the social fabric, from civil society to the justice system. Both concepts can further inspire resistance and transformative practices that empower moral agents striving to promote the common good.[140]

Re-appropriating Human-centered Labor

In *Laborem Exercens* (LE), John Paul II's encyclical addressing human labor, the Pope recognizes how work, a "fundamental dimension" of human existence, characterizes Jesus's incarnation (nos. 26–27) and expresses human dignity, ingenuity, and creativity in the history of humankind, while human beings cooperate with God's creative action in creation (nos. 6, 25).[141] Human work, however, implies "toil

---

[136] See Paul Rabinow and Nikolas Rose, "Biopower Today," *BioSocieties* 1 (2006): 195–217.

[137] Arun, "AI and the Global South," 597–98. Aadhaar also targets other vulnerable people—i.e., undocumented Bangladeshi migrant workers—by making it easier to identify and deport them. See also Castelvecchi, "Beating Biometric Bias," 349.

[138] Arun, "AI and the Global South," 598.

[139] Arun, "AI and the Global South," 598.

[140] See Julio L. Martinez, S.J., "Rivisitare il Bene Comune nell'Era Digitale," *La Civiltà Cattolica* II, no. 4078 (2020): 328–41. See also Michael J. Sandel, *The Tyranny of Merit: What's Become of the Common Good?* (New York: Farrar, Straus and Giroux, 2020). I am grateful to Gustavo Monzon, SJ, for this last reference.

[141] See also Patricia A. Lamoureux, "Commentary on *Laborem Exercens* (*On Human Work*)," in *Modern Catholic Social Teaching: Commentaries and Interpretations*, ed.

and suffering, and also … the harm and injustice which penetrate deeply into social life within individual nations and on the international level" (no. 1). For John Paul II, to reflect on work means stressing the dignity of workers, avoiding commodification and inhuman working conditions, and promoting solidarity among workers (no. 8).[142] Work is good for humankind, because it allows human beings to collaborate with God in creative ways. Human labor allows personal and social flourishing as well as human realization.

At the same time, the Pope's approach acknowledges the complexity of working contexts darkened by the evil of exploitation, abuse, forced migration, "the lack of adequate professional training and of proper equipment, the spread of a certain individualism, and also *objectively unjust situations*" (no. 21). Moreover, "human work is *a key,* probably *the essential key,* to the whole social question" (no. 3). Paying attention to the dignity of work, workers, working conditions, and diversified social contexts is an urgent ethical task.[143]

While technological developments should contribute to the humanization of work, for John Paul II "in some instances, technology can cease to be man's ally and become almost his enemy, as when the mechanization of work 'supplants' him, taking away all personal satisfaction and the incentive to creativity and responsibility, when it deprives many workers of their previous employment, or when, through exalting the machine, it reduces man to the status of its slave" [*sic*] (no. 5).

Because the person is "*the primary basis of the value of work*" (no. 6), it is necessary to address what hinders experiencing work as an essential dimension of human dignity, any working condition that harm workers, and the lack of access to work. For the Pope, the Catholic Church should be firmly committed to caring for the poor, being truly the "Church of the poor," aware that:

> The "poor" appear under various forms; they appear in various places and at various times; in many cases they appear as a *result of the violation of the dignity of human work:* either because the opportunities for human work are limited as a result of the scourge of unemployment, or because a low value is put on work and the rights that flow

K. R. Himes, L. S. Cahill, C. E. Curran, D. Hollenbach, and T. A. Shannon, 2nd ed. (Washington, DC: Georgetown University Press, 2018), 403.

[142] On treating human beings as object and "*instrument of production*," see *Laborem Exercens*, no. 7 (emphasis in original).

[143] See Christine Firer Hinze, *Glass Ceilings and Dirt Floors: Women, Work, and the Global Economy*, 2014 Madeleva Lecture in Spirituality (New York: Paulist, 2015); Christine Firer Hinze, *Radical Sufficiency: Work, Livelihood, and a US Catholic Economic Ethic*, Moral Traditions (Washington, DC: Georgetown University Press, 2021).

> from it, especially the right to a just wage and to the personal security of the worker and his or her family. (no. 8)

In the currently dominant capitalist context, repeatedly John Paul II affirms his personalist approach attentive to the social and productive context by stressing that "*the principle of the priority of labour* over capital is a postulate of the order of social morality" (no. 15)[144] that relies on reaffirming and implementing the rights of workers (nos. 16–23) and promoting education (no. 18).

In her commentary on LE, Patricia Lamoureux engages the encyclical's theological anthropology centered on the preeminence of the subjective dimension of work, the priority of labor over capital, workers' rights, and the spirituality of work.[145] In her assessment, "*Laborem Exercens* provides a good foundation and several building blocks for developing an ethic of discipleship in the workplace. The challenge for the future is to construct an edifice that more closely reflects the reign of God, one that promotes justice for workers, fosters solidarity, and enables workers to become virtuous and self-determining."[146] However, "An ethic of human labor requires more attention to social sin and structures than the encyclical provides."[147] A careful and comprehensive view of work able to address the current changes fostered, among others, by implementing AI technology, should engage "social structures that contribute to work that is meaningless or dehumanizing."[148] As she writes, "The challenge is to create the conditions that make it possible to offer work that satisfies the requirement of self-realization and that enables participation in the workplace."[149] Globalization and technological progress amplify this challenge, as Pope Benedict XVI, Pope Francis, and various theologians have stressed.[150]

---

[144] See also nos. 12–14.

[145] Lamoureux, "Commentary on *Laborem Exercens*," 408–18.

[146] Lamoureux, "Commentary on *Laborem Exercens*," 420. See also Christine Firer Hinze, "Women, Families, and the Legacy of *Laborem Exercens*: An Unfinished Agenda," *Journal of Catholic Social Thought* 6, no. 1 (2009): 63–92.

[147] Lamoureux, "Commentary on *Laborem Exercens*," 420. On social structures, see Daniel J. Daly, *The Structures of Virtue and Vice*, Moral Traditions (Washington, DC: Georgetown University Press, 2021); Daniel K. Finn, ed., *Moral Agency within Social Structures and Culture: A Primer on Critical Realism for Christian Ethics* (Washington, DC: Georgetown University Press, 2020); Daniel K. Finn, *Consumer Ethics in a Global Economy: How Buying Here Causes Injustice There*, Moral Traditions (Washington, DC: Georgetown University Press, 2019), 61–76.

[148] Lamoureux, "Commentary on *Laborem Exercens*," 420. See also David L. Gregory, "*Laborem Exercens*'s Prescient Critique of Technology," *Journal of Catholic Social Thought* 6, no. 1 (2009): 113–31.

[149] Lamoureux, "Commentary on *Laborem Exercens*," 421.

[150] As examples, see Ilsup Ahn, "The Globalization of Labor and the Limits of Sovereignty: Immigration and the Politics of Forgiveness," *Political Theology* 19, no. 3 (2018): 193–210; *Caritas in Veritate*; Bernard Quintard, "De *Laborem Exercens* à *Caritas in Veritate*," *Bulletin de littérature ecclésiastique* 111, no. 1 (2010): 31–44;

John Paul II's vision of work is far from being realized. Changes caused by AI further compel moral agents and civil society to strive for realizing such a vision, with the personal and social flourishing that it encompasses. As Pope Francis reminds us,

> We were created with a vocation to work. The goal should not be that technological progress increasingly replace human work, for this would be detrimental to humanity. Work is a necessity, part of the meaning of life on this earth, a path to growth, human development, and personal fulfilment. Helping the poor financially must always be a provisional solution in the face of pressing needs. The broader objective should always be to allow them a dignified life through work. Yet the orientation of the economy has favored a kind of technological progress in which the costs of production are reduced by laying off workers and replacing them with machines. This is yet another way in which we can end up working against ourselves. (*Laudato Si'*, no. 128)[151]

CONCLUSION

AI could contribute to promoting the common good of humankind and of the planet. To facilitate this goal, while the current ethical agenda generally proposes principles, further ethical integrations are possible.[152] First, the discernment required to address the tension between an ethic of control and an ethic of risk stresses the importance assigned to the moral agent as well as a dynamic understanding of agency. Such an approach seems to be appropriate to reflect critically on the possible beneficial uses of facial recognition technology in diverse social contexts, while avoiding biased forms of control and engaging in carefully evaluated uses.

Second, deploying AI within the legal and judicial system could benefit from a critical reading of structural dimensions by examining power dynamics centered on human bodies. Revisiting the notions of biopower and biopolitics to stress both their deconstructive and constructive components could guide in identifying racially-, gender-, and class-biased abuses harming individuals and curtailing the integrity of

---

Diarmuid Martin, "Catholic Social Teaching and Human Work: The 25th Anniversary of *Laborem Exercens*," *Journal of Catholic Social Thought* 6, no. 1 (2009): 5–17; John A. Coleman, "Pope Francis on the Dignity of Labor," *America,* November 23, 2013, www.americamagazine.org/faith/2013/11/20/pope-francis-dignity-labor; Francis, "Address to Delegates from the Italian Confederation of Workers' Unions (CISL)," June 28, 2017, www.vatican.va/content/francesco/en/speeches/2017/june/documents/papa-francesco_20170628_delegati-cisl.html.

[151] See also Francis, "Address to Delegates."

[152] On AI in healthcare settings and promoting virtues, see Andrea Vicini, SJ, "Artificial Intelligence in Healthcare: Bioethical Challenges and Approaches," *Asian Horizons* 14, no. 3 (2020): 615–27.

the justice system and helping to truly promote justice equally for all citizens.

Third, the transformations AI is progressively introducing in hiring, production, marketing, and workplaces should not harm workers by creating new forms of exclusion, marginalization, abuse, and unemployment. It is urgent to reaffirm the centrality of the person, promote the quality of working conditions, stress the importance of training, converting, enriching, and integrating the workers' skills, together with fostering strong solidarity among workers and in society. These are essential and reachable characteristics of a flourishing marketplace. They could be pursued in innovative ways as an expression of human ingenuity and moral imagination.[153]

Across the planet, colleges and universities have the important role of educating current and future generations by empowering them to make positive contributions in shaping the technological development of AI in the social fabric. Projects and initiatives that foster creative innovation—like human-centered engineering—could lead to developing AI technology in ways that allow to use it for promoting what is good and just: from law enforcement to education, entrepreneurship to the job market. M

Andrea Vicini, SJ (MD, PhD, STD), is Michael P. Walsh Professor of Bioethics and Professor of Theological Ethics in the Boston College Theology Department. Recent publications include two co-edited volumes—*Ethics of Global Public Health: Climate Change, Pollution, and the Health of the Poor* (The Journal of Moral Theology/Wipf and Stock, 2021); *Reimagining the Moral Life: On Lisa Sowle Cahill's Contributions to Christian Ethics* (Orbis Press, 2020)—and two articles—"Artificial Intelligence in Healthcare: Bioethical Challenges and Approaches," *Asian Horizons* 14, no. 3 (2020): 615–27 and "COVID-19: A Crisis and a Tragedy—What's Next?," *Theological Studies* 82, no. 1 (2021): 116–37.

[153] See Patricia H. Werhane, *Moral Imagination and Management Decision-Making*, Ruffin Series in Business Ethics (London: Oxford University Press, 1999). I am grateful to Federico Cinocca for this reference. See also Laura Boella, *Il Coraggio dell'Etica: Per una Nuova Immaginazione Morale* (Milano: Raffaello Cortina, 2012).

Journal of Moral Theology, Vol.11, Special Issue 1 (2022): 70–86

Can Lethal Autonomous Weapons Be Just?

Noreen Herzfeld

In 2018 the United States Department of Defense (DoD) created a new Joint Artificial Intelligence Center to study the adoption of AI by the military. Their strategy, outlined in a document entitled "Harnessing AI to Advance Our Security and Prosperity," proposes to accelerate the adoption of AI by fostering "a culture of experimentation and calculated risk taking," noting that AI will soon "change the character of the future battlefield and the pace of threats we must face."[1] The report cautions that Russia and China are investing in AI for military purposes. While it is the DoD's intention to keep up in the AI arms race, the report states that we will "undertake research and adopt policies as necessary to ensure that AI systems are used responsibly and ethically."[2] How will we determine what is responsible or ethical in this new technological age? Just as the advent of nuclear weapons caused twentieth-century theologians to reevaluate the justness of war, the advent of lethal autonomous weapon systems (LAWS) presents an even more urgent call to twenty-first century theologians to do likewise. For while nuclear weapons required human decision-makers, autonomous systems may not. While nuclear bombs are extremely costly and difficult to produce, intelligent weapons are becoming smaller and cheaper and can be used across the spectrum of conflict, from the highest to the lowest level.

Weapons of war have long had a certain degree of autonomy. Heat seeking missiles can change their course. Defensive systems include automatic modes that target inbound projectiles independent of any human decision.[3] With its potential to bring autonomy to a new level, AI forces us to consider just how independent we want our weapons to be and what difference this heightened autonomy makes to the ethical conduct of war. While bombs, land mines, missiles, and drones do not always involve a direct human decision as to whom they target, nor when exactly they wreak their havoc, they do not make decisions. They cannot decide *not* to explode when triggered. Nor can they target

[1] US Department of Defense, "Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity," (2018), §4, media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.
[2] US Department of Defense, "Artificial Intelligence Strategy."
[3] See for example en.wikipedia.org/wiki/Phalanx_CIWS.

with any degree of specificity. A landmine targets whoever steps on it; a bomb maims or kills whoever is within range. The use of artificially intelligent weapons, especially when combined with capabilities such as facial recognition, inaugurates a new era in weaponry, one which differs from what has preceded it in kind rather than merely degree. Autonomous weapons take over not just the physical, but many of the mental decisions of the battlefield. Just as the necessity for the physical presence of soldiers limited the destruction of war prior to the twentieth century, so the mental limitations of human decision making have continued to function as a limiting principle. Autonomous weapons risk moving humankind into an era of warfare that moves with unprecedented speed, precision, and unexpected consequences.

Ideally, such weapons should be banned, and both individuals and international bodies have indeed called for such a ban. However, given the number of countries already engaging in their development, and the advantages possession of autonomous weapons could confer, an international ban is extremely unlikely. Lacking such a ban, we must at least examine the challenges such weapons will bring to the initiation, execution, and ending of wars and develop procedures and guidelines that ensure these weapons are used in a restrained and just manner. The tradition of just war theory, an amorphous set of rules and justifications rooted in Christian thought and consisting of "articulated norms, customs, professional codes, legal precepts, religious and philosophical principles, and reciprocal arrangements that shape our judgments of military conduct," has provided justification for both when to go to war and acceptable conduct within a war in the past.[4] These precepts illuminate some of the ethical quandaries of warfare that LAWS will exacerbate and highlight areas in which we will need to develop guidelines and policies we currently lack.

Traditional Principles of Just Warfare

There is no one set of principles of just warfare. We find an early rule of conduct for warfare in Deuteronomy 20:19–20 which forbids cutting down an enemy's fruit trees during a siege. This verse counsels both restraint against needless environmental destruction while also hinting at a principle of distinction, for destruction of an enemy's means of producing food harms combatants and noncombatants alike. A Christian tradition of just warfare can be traced back to the pre-Christian thought of Aristotle's *Politics* and expanded upon by Augustine in *The City of God* and *Against Faustus the Manichaean*. It is, however, Thomas Aquinas, in his *Summa Theologiae*, who first lays out the general outline of what we now regard as traditional just war

[4] Michael Walzer, *Just and Unjust Wars: A Moral Argument with Historical Illustrations*, fifth edition (New York: Basic Books, 2015), 44.

theory. These principles were adjusted and universalized to include conduct toward non-Christians by later Scholastics including Vitoria, Suarez, Grotius and Wolff[5] in light of European colonial expansion, and further rethought in the twentieth century by Walzer, Nagel, Norman, and others in light of the changing face of warfare brought about by the atomic bomb and rise of international terrorism.[6] The current field of just war theory is contentious, in that how a state should regulate war and what an individual may be morally obligated to do can conflict. Autonomous weapons add to this confusion by introducing a new potential actor beyond the state and individual, namely, the weapon itself.

Michael Walzer, in his highly regarded *Just and Unjust Wars: A Moral Argument with Historical Illustrations*, lays out a set of traditional principles for when it is permissible to fight and how to engage ethically in battle. *Jus ad bellum*, when it is justifiable to fight, includes fighting for a just cause, either to resist attack, protect innocent life in imminent danger or, as phrased by the US Catholic bishops, to "correct a grave, public evil" such as genocide or a massive violation of a group's basic human rights. Hostilities should be declared by a proper authority, be a last resort, fought only for the purpose that initiated it, have a reasonable chance of success, and be proportional, in that the anticipated benefits of waging war outweigh the expected harm. The calculation of the costs of war should include those paid by one's enemy as well as one's own, both tangible and intangible. These principles act together to militate against going to war, for as Pope Pius XII stated on the eve of World War II, "Nothing is to be lost with peace; everything can be lost with war."[7]

Jus in bello, or proper conduct once engaged in hostilities, includes distinguishing between combatants and non-combatants, limiting the level of force to the minimum necessary to attain one's ends, and acting solely with the intention of righting the wrong. Acts of vengeance and indiscriminate violence are forbidden as are intrinsically evil methods, such as mass rape, forcing enemy combatants to fight against

---

[5] See Francisco de Vitoria, *Relectiones theologicae* (Lyons: Apud Jacobum Boyerium, 1557); Francisco Suárez, *De triplici virtute theologica* (London: Clarendon, 1944); Hugo Grotius, *The Rights of War and Peace: Including the Law of Nature and of Nations* (New York: M. W. Dunne, 1901); Christian Freiherr von Wolff, *The Law of Nations Treated According to the Scientific Method* (Carmel, IN: Liberty Fund, 2017).

[6] Michael Walzer, *Just and Unjust Wars;* Thomas Nagel, "War and Massacre," *Philosophy & Public Affairs* 1, no. 2 (1972): 123–44; Richard Norman, *Ethics, Killing, and War* (Cambridge: Cambridge University Press, 1995); G. E. M. Anscombe, "War and Murder," in *Nuclear Weapons: A Catholic Response*, ed. Walter Stein (London: Sheed and Ward, 1961), 44–52.

[7] Quoted by Rev. Diarmuid Martin, United States Council of Catholic Bishops, "Theological and Moral Perspectives on Today's Challenge of Peace" (2011), www.usccb.org/issues-and-action/human-life-and-dignity/september-11/theological-and-moral-perspectives-on-todays-challenge-of-peace.cfm.

or betray their own side, or the use of weapons whose effects are uncontrollable, such as chemical or biological weapons. Civilians are to be safeguarded as much as possible.

Later scholars, such as Canadian philosopher Brian Orend and ethicists Mark Allman and Tobias Winright, have added to these two categories principles of conduct following the cessation of hostilities (*jus post bellum*).[8] After the cessation of hostilities, the victorious party has the responsibility to lay down arms, enter into relevant treaties, and remove soldiers and weapons from the field of battle. They should also aid in the political and economic reconstruction of the defeated community or state. Both parties should be held responsible for war crimes or atrocities committed during hostilities and the victor is responsible for seeing that suitable restitution or reparations are made, especially in the case of genocide.

A final category, *jus ad vim* (introduced by Michael Walzer in the fourth edition of his seminal text), examines use of force in conditions falling short of full warfare. This category remains controversial due to the difficulty in determining where the line falls between acts just short of warfare and acts of terrorism, criminality, or policing. Many countries have seen the adoption of military-grade hardware by police departments and civilian branches of government. Autonomous weapons might in the future be used domestically in efforts to fight crime or, more sinisterly, to target political opponents, and internationally in efforts to stem terrorism or combat drug and weapons markets. LAWS hold the potential to make assassination or small targeted strikes easier and hence, more likely. Such uses demand a consideration of their potential to escalate a hostile situation and of their effect on the hearts and minds of a populace.

The rapid rise of new technologies of warfare in the twentieth and twenty-first centuries has sparked a commensurate rise in interest in and public debate of the just-war tradition. These precepts have become a part of political deliberations on the use of force and in military training in several countries. The US Council of Catholic bishops sees them as not just "a set of ideas, but as a system of effective social constraints on the use of force." Their application is neither straightforward nor easy, but for the bishops, the "increasing violence of our society, its growing insensitivity to the sacredness of life and the glorification of the technology of destruction in popular culture" call for their use. The speed, precision, and lethality of modern weapons, all characteristics enhanced by AI, make it all the more important that our

[8] See Brian Orend, "Jus Post Bellum," *Journal of Social Philosophy* 31 (2000): 117–37; Mark Allman and Tobias Winright, *After the Smoke Clears: The Just War Tradition and Post War Justice* (Maryknoll, NY: Orbis, 2010).

decisions regarding the use of lethal force at least attempt to pass the "hard tests set by the just-war tradition."[9]

LETHAL AUTONOMOUS WEAPONS AND THE "HARD TESTS" OF JUST WAR THEORY

Computer assisted weapons of war exist along a scale of autonomy. At one end we have traditional "fire and forget" weapons, such as some guided missiles where a human operator selects a target and launches the missile, which then uses sensors and algorithms to complete the task. These clearly depend on a human operator being "in the loop." US Department of Defense Directive 3000.09 notes two other levels of autonomy. "Human on the loop" are weapons that act autonomously but under human supervision, where an operator can monitor the weapon and halt or alter its engagement. Fully autonomous weapons, where the human is "out of the loop" are "weapon system[s] that, once activated, can select and engage targets without further intervention by a human operator."[10] Philosopher Robert Sparrow adds that, to be seen as truly autonomous, such a system must be sufficiently complex, "such that, even when it is functioning perfectly, there remains some uncertainty about which targets it will attack and why."[11]

Lethal autonomous weapon systems have moved in the past decade from the semi-autonomy of "human in the loop" to being able to autonomously identify a target and decide, without synchronous human control, whether to attack or destroy it. As such, they potentially have not only a high degree of agency, acting without the direct control of another to achieve a chosen result but, since their targets are often human beings, they act as moral agents, making judgments and decisions that carry the potential for life or death. The advent of flight inaugurated a new era of warfare, releasing armies from physical presence on the field of battle. Fully autonomous weapons will inaugurate a third era, releasing soldiers from the mental decisions of the battlefield as well. According to legal scholar Rebecca Crootof, their "capacity for self-determined action makes them uniquely effective and uniquely unpredictable."[12] Former USAF Major General Robert Latiff

[9] National Conference of Catholic Bishops, *The Harvest of Justice is Sown in Peace* (Washington, DC: US Catholic Conference, 1994), C. The Centrality of Conscience.
[10] Congressional Research Service, "Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems," *In Focus*, December 1, 2020, fas.org/sgp/crs/natsec/IF11150.pdf.
[11] Robert Sparrow, "Killer Robots," *Journal of Applied Philosophy* 24, no. 1 (2007): 64.
[12] Rebecca Crootof, "War Torts: Accountability for Autonomous Weapons," *University of Pennsylvania Law Review* 164, no. 6 (2016): 1349.

has called this transformation the crossing of a new "moral Rubicon."[13]

The urgency of considering the ethics of LAWS is exacerbated by their evolution in the last few decades from large and costly systems to weapons that are, in the words of Marine Col. James Jenkins, "small, smart, cheap, and abundant."[14] An early semi-autonomous system, the AEGIS naval air defense has increased in autonomy over the last fifty years. Currently it is able to search in the air, on the surface, and underwater, track and guide missiles, and decide autonomously when and where to fire. It can function both fully autonomously or in "human on the loop" mode with operators having the option to override its decisions.[15] It is a big and costly system and currently only being updated by the US and Japan. The HAROP loitering missile, a smaller and more autonomous system developed by Israel and first used by Turkey in 2005, can once launched be controlled either via a two-way data link for "human in the loop" operation or programmed to autonomously recognize and attack high-value targets.[16] Even smaller, the Kargu-2 is a 15-pound multi-copter drone that can be controlled directly or operate autonomously to track, identify (via facial recognition), and engage targets. These drones can work autonomously in swarms of up to 20, either with a human operated drone leading the swarm or fully independently. Turkey has ordered 500 of these for military surveillance and possible attack capabilities.[17] Unlike missiles, drones can be sent to hunt down enemy targets and return, weapons unspent, if none are found. Large swarms of drones can be launched, and if only a small percentage find their target, you still have a win. At least sixteen countries possess armed drones. So far, they operate with humans in the loop, but this could easily change as facial recognition and AI decision making improve. Unlike the nuclear club, limited to a handful of nations, LAWS will proliferate much more easily and widely.

These and other autonomous weapons present military commanders with a variety of incentives for use. They can process vast amounts of data and operate at speeds and levels of precision far beyond human

---

[13] Robert Latiff and Patrick McCloskey, "With Drone Warfare, America Approaches the Robo-Rubicon," *Wall Street Journal*, March 14, 2013, www.wsj.com/news/articles/SB10001424127887324128504578346333246145590.

[14] Jon Harper, "Navy, Marine Corps Officials Worried about Cost-Effectiveness of Unmanned Systems," *National Defense*, April 5, 2017, www.nationaldefensemagazine.org/articles/2017/4/5/navy-marine-corps-officials-worried-about-costeffectiveness-of-unmanned-systems.

[15] "AEGIS Weapon System," *US Navy Fact File*, www.navy.mil/navydata/fact_display.asp?cid=2100&tid=200&ct=2.

[16] "HAROP Loitering Munitions System," www.iai.co.il/p/harop.

[17] David Hambling, "Turkish Military to Receive 500 Swarming Kamikaze Drones," *Forbes*, June 17, 2020. www.forbes.com/sites/davidhambling/2020/06/17/turkish-military-to-receive-500-swarming-kamikaze-drones/#4887e2c5251a.

capabilities, including making rapid decisions in changing circumstances. They can operate in harsh and difficult environments, such as underwater. They are less expensive than human troops and can work long hours without tiring. They can carry out orders with fewer mistakes. Most important, they keep soldiers out of physically and psychologically dangerous or deadly environments. However, these advantages do not come without costs. Just as twentieth-century ethicists and theologians were forced to reevaluate the justness of war in the light of nuclear weapons, so now must we reevaluate the morality of war in light of autonomous weapons. In what ways does the advent of these weapons affect our decisions on when to wage war, how to wage war, and who is responsible for the acts of war? We will examine one question raised by each category of just war theory to provide a brief and partial answer to this question.

Jus ad Bellum: Would LAWS Make War Too Easy?

Theologian Brian Stiltner argues that LAWS could make war too easy. He recalls an episode of the original *Star Trek* entitled "A Taste of Armageddon," in which two planets, Eminar and Vendikar, have completely computerized warfare. Attacks are simulated by computers and those unlucky enough to have been "victims" of the simulated attack are required to report to disintegration chambers. Both planets claim to have found a way of maintaining their infrastructure despite being at war, removing the brutality and destruction. Captain Kirk destroys one planet's computers, stating that war is inherently brutal and messy, which keeps us from going to war lightly or perpetuating it too long.[18]

Kenneth Payne, British scholar of international affairs, argues that like the computers on Eminar and Vendikar, LAWS will not only remove too many of the psychological barriers to war, but will further privilege offense over defense.[19] Consider an attack by a swarm of drones such as the Kargu-2. They can attack *en masse* and then rapidly disperse, leaving little target for a counterattack. Autonomous weapons can also function as "suicide bombers" with no loss of life for the attacker. The direction of the attack can be various and loitering missiles can bide their time or come and go in unpredictable ways. The attacking AI will be expert in observing and analyzing terrain, removing that advantage from the defender, and will be unaffected by fatigue, mental strain, or emotional compunctions.

---

[18] See Brian Stiltner, "A Taste of Armageddon: When Warring is Done by Drones and Robots," in *Can War Be Just in the 21st Century?*, eds. Tobias Winright and Laurie Johnston (Maryknoll, NY: Orbis, 2015).

[19] Kenneth Payne, "Artificial Intelligence: A Revolution in Strategic Affairs?," *Vex Machina*, September 18, 2018, www.tandfonline.com/doi/full/10.1080/00396338.2018.1518374.

Second, use of LAWS removes the constraint of soldiers' lives being put at risk, significantly lowering the cost of an attack. As an editorial in *The Economist* points out, "A president who sends someone's son or daughter into battle has to justify it publicly, as does the congress responsible for appropriations and a declaration of war. But if no one has children in danger, is it a war?"[20] Swarms of intelligent drones would moreover add little material cost to an attack. LAWS thus prove to be a serious risk to the just war precept of last resort. They privilege offense, remove the psychological barrier to putting "boots on the ground," and could significantly lower the cost of an attack, making war seem a more desirable option than it would have been if waged with conventional soldiers and weapons.

Jus in Bello: Can a Robot Act Ethically?

Once at war, the principles of *jus in bello* demand that one act with restraint, that one refrains from gratuitous killing of civilians and excess destruction, and that soldiers conduct themselves with virtue and propriety. Georgia Tech roboticist Ron Arkin has argued that LAWS have the potential to act more virtuously than humans. Arkin cites a report from the Surgeon General's Office assessing the battlefield ethics of US soldiers and marines in which ten percent reported mistreating noncombatants and roughly thirty percent reported facing ethical situations in which they did not know how to respond.[21] Soldiers, under pressure, react emotionally: "Fear and hysteria are always latent in combat, often real, and they press us toward fearful measures and criminal behavior."[22] An analysis of civilian casualties in the second Iraqi war found that most were either the result of ethnic cleansing or caused by indiscriminate suppressive fire between sides. LAWS have no emotions, so will not react out of panic or vengeance. Similarly, they have no need for self-protection. They would follow orders more exactly and can integrate information regarding a changing battle scenario faster before responding with lethal force, thus acting with more precision and fewer mistakes. Arkin believes AIs could better discriminate between combatants and noncombatants, thus committing fewer war crimes and reducing civilian casualties.[23]

Not everyone agrees. John Kaag and Whitley Kaufman argue that moral judgment is inherently ambiguous. Were the laws of war reducible to a set of simple rules, "It is likely that we would have discovered

---

[20] "Drones and Democracy," *The Economist*, October 1, 2010, www.economist.com/babbage/2010/10/01/drones-and-democracy.

[21] Ronald Arkin, "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture," *Technical Report GIT-GVU-07-11*, www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf.

[22] Walzer, *Just and Unjust Wars,* 251.

[23] Arkin, "Governing Lethal Behavior," 6–8.

many or most of these rules long ago."[24] Programming a robot to discriminate between a combatant and a civilian might be easy enough, using facial recognition, in the case of an individual assassination but remarkably difficult in the general context of a counterinsurgency. The US has used "signature strikes" in Pakistan and Afghanistan, authorizing the use of force against any who fit certain behavioral profiles, such as transporting weapons or congregating as large groups of young men. This has, unfortunately, resulted in the targeting of wedding parties in a part of the world where the shooting of rifles is part of the traditional celebration and gender exclusivity separates male and female wedding participants.[25] Discrimination requires a high level of context sensitivity, one that would be complex to program.

Arkin notes that soldiers often violate the principle of right intention, acting out of fear or anger. However, the emotions and consciences of human soldiers also act, at times, as a check on unjustifiable commands or illegal orders. Could a future autonomous weapon have a conscience or true moral agency? Current AIs do exactly what we tell them to do, even when their instructions or the sequence of their learning might be so complex that we cannot anticipate the result. A machine with moral agency would have a further ability to reason independently and unpredictably change course, should it consider the actions it is programmed to take unethical or in violation of an overarching value or intention.[26] Michael and Susan Anderson go a step further. They consider a robot or program to be a moral agent if it fits three criteria. First, it is not "under the direct control of any other agent or user." Second, its interaction with its environment is "seemingly deliberate and calculated." Third, it fulfills "some social role that carries with it some assumed responsibilities."[27] This third criterion points to the relational nature of what we call conscience. We learn our social responsibilities gradually from parents, peers, our faith traditions, and, for a soldier, from his or her fellow soldiers, commanding officers, and basic training. Most soldiers report that the greatest motivating factor for their actions on the battlefield is their sense of solidarity with and responsibility for their fellow soldiers.

Could such social awareness—indeed, obligation—be instilled in AI? Isaac Asimov envisioned such a need and developed what we now call the Three Laws of Robotics:

---

[24] John Kaag and Whiltel Kaufman, "Military Frameworks: Technological Know-how and the Legitimization of Warfare," *Cambridge Review of International Affairs* 22, no. 4 (2009): 601.

[25] For more on US use of drones and signature strikes, see Kenneth R. Himes, *Drones and the Ethics of Targeted Killing* (Lanham, MD: Rowman and Littlefield, 2015).

[26] Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (New York: W. H. Freeman, 1976), 74.

[27] Michael Anderson and Susan Leigh Anderson, *Machine Ethics* (Cambridge: Cambridge University Press, 2011), 158.

– A robot may not injure a human being or, through inaction, allow a human being to come to harm.
– A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
– A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

While these laws sound good on first encounter, Asimov's two collections of short stories—*I, Robot* and *The Rest of the Robots*—present a thorough exploration of the many situations in which these laws are destined to fail, as they are overly broad, lack contextual adaptation, and can be self-conflicting. The robot is often helpless to act, hardly what any military would want in an autonomous weapon. Instead, we wish to empower such weapons to adapt to changing conditions and make "smart" choices. But on what basis? Human soldiers bring years of social experience that hones awareness of both their role and their responsibilities *vis-à-vis* other humans. As yet we have no idea how to instill this into a machine.

Expecting LAWS to engage in just warfare assumes not only that its precepts are codifiable or learnable by AI, but that they are what would actually be programmed or taught. Moreover, an AI's original programming might be altered or hacked by a bad actor. Should autonomous weapons become "small, smart, cheap, and abundant," it will be hard to keep them out of the hands of terrorists and non-state actors, who would not have the same interest in following the rules of warfare. Even lacking these scenarios, how likely is a military to prioritize ethics over victory? I fear that a nod to ethical principles could easily degrade into mere "window dressing" for the public while the true goal in programming LAWS would be, and many would argue rightly so, to win at all costs.

Finally, sometimes moral behavior means breaking the rules. Most disputes between Jesus and the Pharisees recounted in the Gospels hinged precisely on Jesus or his disciples breaking a rule or religious convention. The spirit of the law does not always match the letter. Former Army Ranger Paul Scharre recounts a situation in the Iraq war in which the Mahdi Militia used a child as a forward observer. US forces did not shoot the child, even though the conventions of war would have allowed it.[28] Would AI be programmed with sufficient nuance to make this judgment call? This is certainly beyond our programming capabilities at this time. Hence, I must disagree with Arkin; for now, it seems unlikely that an autonomous weapon would behave more justly than a human being.

[28] Michael Anderson and Susan Leigh Anderson, *Machine Ethics*, 600.

Jus post Bellum: Who Was Responsible?

In *The City of God*, Augustine notes that the goal of every war is a final state of "peaceful order": "It is an established fact that peace is the desired end of war. For every man is in quest of peace, even in waging war, whereas no one is in quest of war when making peace."[29] While not a part of traditional just war theory, the aftermath of war may be as important in reestablishing order and right relationship between the warring parties as the war itself. One might think LAWS, being weapons, would need no further consideration once the shooting stops. However, one necessary process in the period immediately following a war is the determination and execution of retributive and/or restorative justice. Despite the best planning injustices, atrocities, accidents, and war crimes will occur. When these are not addressed, grievances may fester.

Eventually, an autonomous weapon will be involved in an accident or atrocity that seriously violates international law or Christian ethics. When this occurs, who is responsible? Recall that one of the stipulations for a weapon to be considered autonomous is that its choices and decisions must carry a certain degree of unpredictability. A computer program that is entirely predictable is completely determined by its programmer. While the actions of an autonomous weapon may be foreseeable in most circumstances, they will not always perform as expected. Indeed, systems that depend on machine learning can be "opaque even to the system's designers."[30] While the designers or programmers carry a certain degree of responsibility for creating such a machine, can they be held responsible for any particular decision?

If not the programmer, can the machine itself be held responsible? Rebecca Crootof argues that there is no sense in this. To call something a war crime it must have been "willfully" committed. At this point, and in the foreseeable future, we cannot say a machine behaves either intentionally or recklessly. Nor would it make any sense to punish a machine that can feel neither emotional nor physical pain. The machine can certainly be decommissioned, but it will feel no sense of responsibility. According to Crootof, "Traditional justifications for individual liability in criminal law—deterrence, retribution, restoration, incapacitation, and rehabilitation—do not map well from human beings to robots."[31] Ultimately, we expect another human to carry the mantle of responsible agent.

That leaves the one who deploys LAWS responsible. Under current military law, a commanding officer can be held indirectly responsible for the actions of those under his or her command if those actions

[29] Augustine, *Concerning the City of God against the Pagans*, trans. Henry Bettenson (London: Penguin, 1984), 866.
[30] Crootof, "War Torts," 1373.
[31] Crootof, "War Torts," 1377.

could have been in any way foreseen or prevented. A panel at Harvard Law School noted the challenge to this raised by LAWS' inherent unpredictability: "Would fully autonomous weapons be predictable enough to provide commanders with the requisite notice of potential risk? Would liability depend on a particular commander's individual understanding of the complexities of programming and autonomy?"[32]

Since none of the other options work, the responsibility to see that LAWS are designed to act justly must therefore remain, for now, with the state that deploys them, as does any responsibility for restorative or retributive justice in the event of a moral breach.[33]

Jus ad Vim: Use of Autonomous Weapons Outside of War

LAWS need not be confined to the battlefield. They might also become a weapon of choice against internal enemies. In a 2015 statement, Amnesty International cited their concern regarding the possible deployment of weapons with facial recognition or selection abilities based on other physical traits against minorities or political targets.[34] In a special report to the Council on Human Rights of the UN, Christof Heyns similarly raised this concern: "On the domestic front, [LAWS] could be used by States to suppress domestic enemies and to terrorize the population at large, suppress demonstrations and fight 'wars' against drugs."[35]

On the foreign front, LAWS could and likely will replace unmanned drones as the weapon of choice in combatting terrorism and effecting assassinations. We have already seen how drones have increased the incidence of attacks within the boundaries of states with whom we are not at war, specifically in Pakistan and Yemen in recent years. Use of LAWS presents the danger of automatically escalating hostilities and, more importantly, hardening the hearts and minds of the civilian populations under anonymous surveillance and ongoing threat of attack. The low costs and lack of danger to one's own soldiers presented by LAWS could easily tip calculations toward, rather than away from, escalating a conflict.

Berkeley AI researcher and activist Stuart Russell notes: "I'm not too worried about vast autonomous swarms of battle tanks…there are far cheaper ways to flatten a city and/or kill all of its inhabitants." Instead, Russell fears exactly those weapons that would be used off the battlefield—cheap, small, lethal drones that could be used by police,

---

[32] Crootof, "War Torts," 1381.
[33] Crootof, "War Torts," 1390.
[34] Amnesty International, "UN: Ban Killer Robots before Their Use in Policing Puts Lives at Risk," April 16, 2015, www.amnesty.org/en/latest/news/2015/04/ban-killer-robots-before-their-use-in-policing-puts-lives-at-risk.
[35] Christof Heyns, "Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions," April 9, 2013, www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47_en.pdf.

terrorists, or indeed anyone with a grudge. "My UAV colleagues tell me they could build a weapon that could go into a building, find an individual, and kill them as a class project."[36] This is a frightening scenario, and perhaps the strongest reason many have for an international ban on such weapons.

SHOULD LAWS BE BANNED?

We have only begun to include LAWS in deliberations on arms control. The International Committee of the Red Cross has published advisory guidance on the use of autonomous weapons, but there are no formal international agreements.[37] War must be waged by a responsible and legitimate authority, at all levels. Robert Sparrow argues that, since no human can ultimately be held responsible for their actions, LAWS are profoundly and irremediably unethical.[38] Several international groups agree. The Campaign to Stop Killer Robots, a coalition of 166 non-governmental organizations in sixty-six countries, has called for a ban on fully autonomous weapons. They maintain that the use of such weapons "crosses a moral threshold" because machines "lack the inherently human characteristics such as compassion that are necessary to make complex ethical choices."[39] They fear that such weapons would lower the barriers against going to war and further shift the burden of warfare onto civilians. Amnesty International has also called for a ban, fearing the human rights implications of such weapons.

Others fear a new arms race. An open letter, signed by such notables as Elon Musk and Stephen Hawking as well as over 4500 other researchers in AI and robotics, published by The Future of Life Institute supports a total ban lest autonomous weapons "become the Kalashnikovs of tomorrow." They note that because autonomous weapons are not nearly as costly or difficult to produce as nuclear weapons it may "only be a matter of time until they appear on the black market and in the hands of terrorists, dictators wishing to better control their populace, warlords wishing to perpetrate ethnic cleansing, etc. Autonomous weapons are ideal for tasks such as assassinations, destabilizing nations, subduing populations and selectively killing a particular

---

[36] Sydney Freedberg, "Genocide Swarms & Assassin Drones: The Case for Banning Lethal AI," *Breaking Defense*, March 8, 2019, breakingdefense.com/2019/03/genocide-swarms-assassin-drones-the-case-for-banning-lethal-ai.

[37] International Committee of the Red Cross, "Autonomous Weapon Systems: Is It Morally Acceptable for a Machine to Make Life and Death Decisions?," April 13, 2015, www.icrc.org/en/document/lethal-autonomousweapons-systems-LAWS.

[38] Sparrow, "Killer Robots," 64.

[39] Campaign to Stop Killer Robots, "Statement to the Informal Discussions on Autonomous Weapon Systems," June 29, 2021, www.stopkillerrobots.org/wp-content/uploads/2021/09/CSKR-Statement-to-the-informal-discussions.docx.pdf.

ethnic group."[40] In light of the work of these groups and many others, in 2018 the European Parliament called upon the United Nations General Assembly to "work towards an international ban on weapon systems that lack human control over the use of force" and "to urgently develop and adopt a common position on autonomous weapon systems."[41]

The Vatican agrees, finding it an affront to human dignity to be killed by a machine that cannot make "intentional, rational, and deliberate decisions from a moral and ethical standpoint."[42] Speaking as Vatican observer to the UN in Geneva, Archbishop Ivan Jurkovic stated that the development of autonomous weapons would provide "the capacity of altering irreversibly the nature of warfare, becoming more detached from human agency, putting in question the humanity of our societies."[43] According to Jurkovic, each new trend in weaponry "contributes to increasing awareness that the cruelty of conflicts must be done away with in order to resolve tensions by dialogue and negotiation."[44]

Former lieutenant colonel David Grossman suggests that killing has, in the past, not come easily to human soldiers. He cites a study conducted by the US Army stating that only 15 to 20 percent of soldiers fired their weapons in combat in World War II. Fewer fired to kill.[45] These percentages rose in subsequent wars; Grossman cites two factors that work together to overcome our resistance to killing one another. The first is dehumanization of the enemy. The more the enemy is seen to be "like us," the harder it is to kill that person. The second is distance from the target. Hans Morgenthau has suggested that the increasing automation of war overcomes both these factors, bringing us close to "push-button war," war that is "anonymously fought by people who have never seen their enemy alive or dead and

---

[40] "Autonomous Weapons: An Open Letter from AI & Robotics Researchers," futureoflife.org/open-letter-autonomous-weapons.

[41] "European Parliament Recommendation to the Council on the 73rd session of the United Nations General Assembly (2018/2040[INI])," www.europarl.europa.eu/doceo/document/A-8-2018-0230_EN.html?redirect#title1.

[42] Liam McIntyre, "Autonomous Weapons Systems Threaten Peace, Says Vatican Official," *Crux*, March 29, 2019, cruxnow.com/vatican/2019/03/autonomous-weapons-systems-threaten-peace-says-vatican-official.

[43] Catholic News Service, "Vatican Official: Prohibit 'Killer Robots' Now before They Become Reality," *The Tablet*, November 28, 2018, www.thetablet.co.uk/news/11072/vatican-official-prohibit-killer-robots-now-before-they-become-reality.

[44] Catholic News Service, "Holy See Renews Appeal to Ban Killer Robots," November 28, 2018, www.catholicnewsagency.com/news/holy-see-renews-appeal-to-ban-killer-robots-74479.

[45] S. L. A. Marshall, *Men against Fire: The Problem of Battle Command* (Norman, OK: University of Oklahoma Press, 2000).

who will never know whom they have killed."[46] Distance and dehumanization go hand in hand. In a 2013 "Resolution against Drone Warfare," the Church of the Brethren noted: "Jesus, as the Word incarnate, came to dwell among us (John 1:14) in order to reconcile humanity to God and bring about peace and healing…. We find the efforts of the United States to distance the act of killing from the site of violence to be in direct conflict to [this] witness of Christ Jesus."[47]

International treaties have banned or limited chemical and nuclear weapons. While such bans have not been honored by all countries and regimes, they have served to keep these weapons out of many national arsenals. While ideal, a total ban on autonomous weapons is unlikely. At the level of the United Nations, it is likely that the US, Russia, and China, each with a vigorous LAWS program, would veto such a ban. The pace of technological development far outstrips that of diplomacy. Lacking a total ban, the rules of just war theory become that much more important. In their recognition that warfare is at times regrettably unavoidable, they provide a system for mitigating war's reach and effects. Scharre suggests that we use just war traditions to establish "rules of the road" for autonomous weapons, rules that would at minimum serve to reduce civilian casualties, militate against escalation, and promote transparency.[48]

CONCLUSION: THE ROLE OF HUMAN JUDGMENT

We have noted that lethal autonomous weapons could make war too easy, act without morals, muddy the attribution of responsibility, and lead to more acts of violence outside the already established norms and conventions of warfare. While the optimal solution would be a ban, that remains unlikely. Thus, we now need to establish new norms and conventions, ideally based on just war traditions, to ensure these weapons are used in a restrained and responsible manner.

This calls for human judgment, which has been implicit but rarely specified in the rules of warfare.[49] As we design and build autonomous weapons there are two directions we can take. The first is to continue designing weapons that wage war in our stead. Alternatively, we can design for human-machine symbiosis, leveraging the distinctive strengths of the computer to work together with human beings "to

[46] Hans Morgenthau, *Politics among Nations* (New York: McGraw-Hill, 2006), 250.
[47] Church of the Brethren Ministry and Mission Board, "Resolution against Drone Warfare," March 2013, www.brethren.org/about/statements/2013-resolution-against-drones.pdf.
[48] Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York: Norton, 2018), 356–57.
[49] Scharre, *Army of None*, 357.

empower, not replace, those who serve."[50] AI should not take over tasks from humans but use differing talents and opportunities to work together to complete those tasks. We must retain a role for human judgment.

Is this, like an international ban on LAWS, also unlikely? While most commanders express a desire for autonomous weapons to have humans in or at least on the loop (having, at minimum, the ability to veto the machine's decisions), how much control can they actually have if decisions in the field are made at a speed humans are unable to follow? The tempo of war has steadily accelerated, increasing dramatically in recent years.[51] LAWS will only further this acceleration. At what point might things move so quickly that we would be forced to cede all decision making to machines? As AI moves from tactical to strategic decision making, this could eviscerate any meaning from the concept of "mission command."[52]

We can still choose not to go down that road. At a recent workshop sponsored by the law faculty at Penn State, I was heartened to hear military commanders, both active and retired, express their personal distaste for LAWS. I join with these commanders in hoping that we choose to never reach the point where LAWS outstrip human commanders's ability to control them. Whether on the field of battle or in the workplace, human dignity depends on our working with our tools rather than letting them supplant us, and this is most important in matters that involve questions of life and death. After watching the first test of a nuclear bomb in 1945, Harry Truman wrote: "Machines are ahead of morals by some centuries, and when morals catch up perhaps there'll be no reason for any of it."[53] International cooperation, treaties, and overwhelming abhorrence have kept nuclear weapons from being used since Truman's time. Let us hope that seventy-five years from now we will have similarly limited the usage of autonomous weapons. It is past time for our morals to catch up with our technology. JM

Noreen Herzfeld is the Nicholas and Bernice Reuter Professor of Science and Religion at St. John's University and the College of St. Benedict and a

[50] Rebecca Slayton, "The Promise and Peril of Artificial Intelligence: A Brief History," *War on the Rocks*, June 8, 2020, warontherocks.com/2020/06/the-promise-and-risks-of-artificial-intelligence-a-brief-history.
[51] T. K. Adams, "Future Warfare and the Decline of Human Decision-Making," *Parameters: US Army War College Quarterly* 2, no. 57 (2001), 57–71.
[52] Kenneth Payne, *Strategy, Evolution, and War: From Apes to Artificial Intelligence* (Georgetown: Georgetown University Press, 2018), 183.
[53] Quoted in Armin Krishnan, *Killer Robots: Legality and Ethicality of Autonomous Weapons* (Farnham: Ashgate, 2009), 167.

research associate with ZRS Koper. She holds degrees in Computer Science and Mathematics from The Pennsylvania State University and a PhD in Theology from The Graduate Theological Union, Berkeley. Herzfeld is the author of *In Our Image: Artificial Intelligence and the Human Spirit* (Fortress, 2002), *Technology and Religion: Remaining Human in a Co-Created World* (Templeton, 2009), *The Limits of Perfection in Technology, Religion, and Science* (Pandora, 2010) and editor of *Religion and the New Technologies* (MDPI, 2017).

Journal of Moral Theology, Vol. 11, Special Issue 1 (2022): 87–111

Artificial Intelligence and the Marginalization of the Poor

Levi Checketts

At the end of Ridley Scott's 1982 film *Blade Runner*, after nearly killing the film's protagonist Deckard, replicant Roy Batty reflects: "Quite an experience to live in fear—that's what it's like to be a slave."[1] The movie's replicants are biologically engineered humans, but the book on which it is based, Philip K. Dick's *Do Androids Dream of Electric Sheep?*, portrays Batty and friends as androids, artificially intelligent robots. Nonetheless, the setting of the story and themes of the film set up an interesting question which, until now, has been radically under-addressed in artificial intelligence (AI) ethics: how can AI inform our understandings of the experience of poverty and our obligation to the poor and oppressed?

This question is of major concern for Catholic social thought in its relation to AI. The so-called "preferential option for the poor," a major pillar of Catholic thought in the modern era, requires us to consider seriously the experience and plight of those who are worst off. While this has often been understood to mean arranging material and economic systems to elevate the poor, deeper reflections—such as those articulated by Latin American liberation theologians—emphasize the need to understand the lived experience of poverty and not just do charitable works. In other words, the issue is not just the problem of wealth inequality but also how one's experience of the world is colored by being poor. Applied to AI, our social inquiries must not only ask whether AI will make some richer and others poorer, but also whether it will better help give authentic voice to the voiceless in our societies.

In this paper I contend that as currently designed, not only will AI systems be unable to articulate the "intelligence" of the poor, but—worse—they will serve to further marginalize the experience and embodied truth of those worst off in our society. The assumptions of what makes something "intelligent" grounded in the aim of artificial intelligence, that is the epistemological models on which AI is built, are rooted in a predominantly bourgeois, Enlightenment-based

[1] Ridley Scott, *Blade Runner*, Warner Bros., 1982.

epistemological model which ignores the ways socioeconomic conditions shape cognition and experience of the world. AI thus represents a version of human intelligence which does not actually correspond to the epistemology of the poor. Worse, as AI becomes a more accepted model of human consciousness, it serves to delegitimize and derogate nonconforming epistemes.

Through this paper, I proceed by first examining the epistemological assumptions of AI research, especially general AI. Following this, I consider how this epistemology conflicts with an epistemology informed by a hermeneutic of poverty. I approach this section by first summarizing some critiques of AI already laid out through leading philosophers and social theorists, and then proposing a hermeneutical framework of the experience of poverty through work done in social theory and liberation theology. Finally, using work done in Science and Technology Studies (STS), I examine the problem of AI "constructing" human intelligence and the potential this has for undermining a social order that takes seriously the experience of the poor.

FEEDBACK SYSTEMS AND COMPUTATIONAL MACHINES

From the beginning, a critical point needs to be made clear: not all AIs are created the same. The term AI is used rather equivocally, in fact, and this is somewhat deliberate. First, when most people think of AI, they envision a computer that is conscious and thinks like we do or, depending on their philosophy of mind, at least can act like it.[2] This view is sometimes called Strong AI, or General AI. Second, however, when the term AI is used by major tech firms like Google, Samsung, Facebook, Apple, and others to sell products, or make big announcements, they often use AI as a synonym for advanced software programs using sophisticated algorithms to make automated decisions in narrow AI. Most often, these are examples of machine learning, adaptive programming, or merely instances of automated assistance. AI is used for IBM's Watson, and DeepMind's AlphaGo, which are highly advanced learning programs rivaling human cognitive abilities in some areas, as well as Alexa and Siri, which are much more akin to Microsoft Word's "Clippy."[3] The reason why a Samsung air conditioner's adaptive setting is called AI is not the same reason why the people at Google's DeepMind call their work AI. Indeed, the people at DeepMind call their work AI because it is—or at least they believe

[2] This idea undergirds much of philosophical discussion of AI in general, from the 1970s until present. Authors such as the late Hubert Dreyfus, John Searle, Nick Bostrom and, most recently, Mark Coekelbergh focus on this aspect of AI as among the most paradigmatic and most philosophically controversial. See, e.g., Mark Coekelbergh, *AI Ethics* (Cambridge, MA: MIT Press, 2020), 11.

[3] Clippy was a Microsoft Office "Assistant" bot that offered suggestions to users, often about formatting, searching for help or optimizing their programs, from 1997 until 2007.

it is—part of the process necessary for achieving General AI.[4] While they recognize that machine learning is not, itself, yet the achievement of a conscious machine, they believe that machine learning and adaptive programming help us see how humans think and how to program computers to do the same.

The push for developing a human-like calculating machine has a surprisingly long genealogy. In 1666, Gottfried Wilhelm Leibniz hypothesized that logical thought was merely a consequence of manipulating inputs and receiving expected outputs, so a machine could be built that housed an "alphabet of human thoughts" and could process human ideas as a mind does.[5] Ada Lovelace, considered by many to be the first programmer, argued in contrast that a machine can never be as intelligent as a human because it does not create original works.[6] However, artificial intelligence work got its applied grounding in the 1940s. The computer pioneer Alan Turing is credited with proposing the first clear vision of AI in terms of human intelligence. He believed digital computers—that is, computing machines with components that are either on or off—were Universal Turing Machines, machines that can simulate any other machine. In his 1950 essay "Computing Machinery and Intelligence," Turing hypothesized that a digital computer can be said to "think" if it can successfully win the "Imitation Game," a test where a judge must choose which responses to questions are from a computer and not a human being, now commonly referred to as the Turing Test.[7] His hypothesis sets the tone and direction of computer engineering generally and AI specifically. To this day, in popular opinion, a computer that successfully passes the Turing Test is, for most purposes, a conscious person.

Also in 1950, Norbert Wiener published *The Human Use of Human Beings*, the foundational text for the field of cybernetics. Within this text, Wiener advanced the ontology that everything is reducible to information. Material is insignificant, and the content of all that is can be understood through its informational context. Indeed, predicting the surge in genomics following Crick, Watson, and Franklin's Nobel Prize-winning work on DNA, Wiener asserted: "We are not the stuff that abides, but patterns that perpetuate themselves."[8] Following this,

---

[4] See "About," *DeepMind*, deepmind.com/about.

[5] Oscar Schwartz, "In the 17th Century, Leibniz Dreamed of a Machine That Could Calculate Ideas," *IEEE Spectrum* (November 4, 2019), spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-the-17th-century-leibniz-dreamed-of-a-machine-that-could-calculate-ideas. See Gottfried Wilhelm Leibniz, *Dissertation on Combinatorial Art*, ed. M. Mugnai, H. van Ruler, trans. M Wilson (Oxford: Oxford University Press, 2020).

[6] Alan Turing, "Computing Machinery and Intelligence," *Mind* 49 (1950): 446.

[7] Turing, "Computing Machinery and Intelligence," 433.

[8] Norbert Wiener, *The Human Use of Human Beings: Cybernetics and Society* (London: Free Association, 1950), 130.

the physical movement of things in space is a question of information and feedback. Cybernetics as a field, then, focuses on feedback to stimuli, often through mechanistic functioning. Wiener and others also believed the human mind was essentially a massive feedback mechanism. Hypothesizing about automated processes, Wiener writes that "the nervous system and the automatic machine are fundamentally alike in that they are devices which make decisions on the basis of decisions they have made in the past"—i.e., a programmed response.[9] Cybernetics, then, reinforced the view of Turing that the human mind is the same in its operation as and can be replicated by a machine. When interpreted by the human brain, a given input of information will yield a predictable and given output. If a human brain can be understood as a sophisticated information processor, then it is conceivable that a sophisticated information processor can replicate the phenomenon of "consciousness" humans seem to possess.

From the 1950s until now, the field of AI research has pursued the dream of replicating human conscious experience. While modern applications of AI are now often directed toward applied types of cognitive activity, such as stock trading, text preservation, or prison bail determinations, all areas of AI work assume the human capacity for any given tasks as the baseline, and many researchers still aim toward the holy grail of General AI. Many obstacles in the process of achieving this have directed contemporary AI researchers into paths the pioneers of the field did not anticipate seventy years ago, such as neural networks, machine learning, and natural language processing, which seek to emulate the more organic way by which human cognition occurs. Nonetheless, no AI has clearly passed either the Turing Test or other proposed human intelligence tests such as the Winograd Schema Test.[10] Each measure of AI's capability is predicated upon an understanding of human intelligence *as primarily a function of calculation and information processing*. Indeed, Ray Kurzweil—one of today's leading artificial general intelligence (AGI) theorists—assumes that the primary uniqueness of human minds is pattern recognition and nothing more.[11]

The measure of AI's success, therefore, is ultimately a measure of how well the program can solve puzzles, perform intellectual tasks, or carry out basic conversations assuming, in each of these, specific metrics for satisfactory or unsatisfactory performance. AI is, for the most

---

[9] Wiener, *The Human Use of Human Beings,* 48.

[10] The Winograd Schema test asks the participant to identify ambiguous pronouns, such as "it" in the sentence, "Mary saw the puppy in the window and wanted *it*." See Ernest Davis, Leora Morgenstern, and Charles Ortiz, "The Winograd Schema Challenge," New York University, cs.nyu.edu/~davise/papers/WinogradSchemas/WS.html.

[11] Ray Kurzweil, *The Singularity is Near: When Humans Transcend Biology* (New York: Viking, 2005), 107.

part, not being trained to do things human beings cannot already do, and those things it is programmed to do tend to be strictly analytic processes. While these may be hallmarks of human distinctiveness in the world, one would hardly call the achievements of AI exhaustive of human intelligence or consciousness. The science fiction depiction of AI, for example, has often hinted at the problem of emotional intelligence, itself only one aspect of the larger depiction of human experience, which includes our senses of wonder and awe, relationality, creativity and aesthetics, moral responsibility, and, most importantly for this work, the interdependent nature of human consciousness and experience of the world. These remain excluded from any meaningful AI research, yet AGI is supposed to replicate human intelligence.

I thus challenge AI's representation of human intelligence as calculation below on two grounds. First, it is dismissive of the varieties of human cognitive experience. In particular, I will consider the disjuncture between AI models of consciousness and the consciousness of the poor, a disjuncture which should be enough to complicate the language and goals of AI. Second, this assumption denies the human dignity of those whose cognitive function is not represented by AI programming goals. If calculation is the model for human cognition, human dignity tends to be contingent upon calculative rationality. This is of special concern for Christians, especially Catholics, whose social obligations prioritize the voice of the voiceless and the experience of those living in the margins of society.

COMPUTER SCIENCE FROM BELOW

The vision of intelligence generally and human intelligence in particular held by AI theorists is one that is highly specific to a particular set of philosophical anthropological beliefs. Hubert Dreyfus notes that the model of epistemology underlying AI theory is one inherently Platonic: human rationality operates on a largely mathematical, formal understanding of the world, with objects corresponding to pure forms and judgments resulting from implicit or explicit calculations.[12] The world is formal, logical, and mathematical. Hava Tirosh-Samuelson also points out latent Cartesianism manifest in the presupposition that intelligence or consciousness is separable from the embodied context. Intelligence is ethereal and programmable into machinery, and embodiment only serves to deceive our rational minds.[13] Noreen Herzfeld takes this further, framing it not merely as a relative problem of philosophical school, but of privileged epistemological context. She notes

[12] Hubert Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (Cambridge, MA: MIT Press, 1994), 67.

[13] Hava Tirosh-Samuelson, "Engaging Transhumanism," in *H±: Transhumanism and Its Critics*, ed. G. Hansell and W. Grassie (Philadelphia: Metanexus Institute, 2011), 44.

that the perspective of intelligence presented by AI—typical of wealthy, able-bodied, heterosexual, white men—excludes many other viable perspectives; intelligence is what fits into the hegemonic models of intelligence in Western society.[14]

The suggestion offered by these authors and others is not that Aristotelian or Leibnizian epistemology will fix the problem, but rather that different contexts of thinking beyond the typical Western philosophical "canon" are necessary. How to get beyond these assumptions is an important question for AI ethics. Dreyfus began such a venture by bringing continental philosophy in as a critique against AI. Using the epistemologies of Ludwig Wittgenstein and Maurice Merleau-Ponty, Dreyfus argued that AI programmed with all information will fail because human beings encounter the world as a *gestalt* and operate in that world based on "rules of thumb" (heuristics) rather than algorithmic procedures.[15] His own thought revolutionized AI and encouraged researchers to pursue machine learning and neural networks rather than massive manual data coding.

Dreyfus's thought does a great deal to "humanize" intelligence, but his context is still as a white man. Critical AI theorists, especially feminist AI theorists, have expanded this vision to include the voices of others. Sherry Turkle and Seymour Papert, for example, found that the gendered assumptions of logical processing and coding were more a product of convention than necessity in computer programming, and that intuitive approaches to computer programming have the same efficacy as procedural approaches.[16] Their work challenges both gendered understandings of logical talent and the dogmatic epistemologies of computer science. Donna Haraway has challenged much in science and technology, including the sterilized "god's eye view" image of science endorsed by white male scientists.[17] She challenges the binary ontologies of roboticists and AI researchers with a hybridized "cyborg" ontology she proposes for feminist approaches to science

---

[14] Noreen L Herzfeld, *In Our Image: Artificial Intelligence and the Human Spirit* (Minneapolis: Augsburg Fortress, 2002), 73. As AI has advanced as a field, it has gone beyond the purview of white Western males to include much research from Asia, especially China. In the US context, however, much of the tech industry is still dominated by white men, though more Asian men now work as software engineers than twenty years ago. The theoretical groundings and pedagogical stylings of the field, however, belie this seeming diversity, and recent problems of "racist" AI illustrate this reality quite clearly.

[15] Dreyfus, *What Computers* Still *Can't Do*, 296.

[16] Sherry Turkle and Seymour Papert, "Epistemological Pluralism: Styles and Voices within the Computer Culture," *Signs* 16, no. 1 (Autumn 1990): 128–57.

[17] Donna Haraway, "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective," *Feminist Studies* 14, no. 3 (Autumn 1988): 575–99.

and technology.[18] Francesca Ferrando's post-human philosophy welcomes the possibility of "artificial" intelligences, but rejects both dualism and the humanist tradition typical of AI epistemology. Artificial intelligence, far from being a replica of human consciousness, should expand our perspectives of what "counts" for intelligence and consciousness.[19] Each of these authors, and many others, assert authentic experiences of the human which are not typified by thought models that support regnant epistemological frameworks in computer science generally and AI especially.

The focus of the rest of this paper is to uplift such a challenge to AI that should be normative for Catholic thinkers—the position of the poor. To frame this challenge, it is important to consider some elements of the lived experience of the poor, with the caveat, as Jon Sobrino notes, that the poor are diverse: poverty exists as disability, sexual discrimination, violence, and other forms of silencing, and looks different across demographics and geographies.[20] Put another way, there is no one poverty, as poverty exists as an element of social marginalization. Something like Kimberlé Williams Crenshaw's thesis of "intersectionality" is key for understanding how poverty is experienced by different people in different ways;[21] the poverty of a disabled peasant widow in El Salvador is different from the poverty of a child factory worker in Nepal, which is also different from the poverty of a low-income black family in the United States.

Despite the rich diversity presented by poverty writ large, the fact remains that nearly all theology is written from non-poor perspectives. Much more needs to be written "from below," by those who have lived in poverty and not merely studied it. I enter this conversation as someone who grew up in first-world poverty, experiencing lack of resources, stunted opportunities, economic contingency, and so forth. Nonetheless, there are important aspects of poverty I never experienced, such as the ways race, gender, disability, and addiction often exacerbate or contextualize the struggles of being poor. The hermeneutic of poverty I outline below reflects my experience but also trends generally among the poor. In a nearly meaningless generalization, I outline in this section poverty as the experience of resource scarcity and insecurity, marginalization and domination. The particular areas I examine in which the normativizing lens of AI

---

[18] Donna Haraway, "A Cyborg Manifesto," in *Simians, Cyborgs, and Women: The Reinvention of Nature* (New York: Routledge, 1991), 149–82.

[19] Francesca Ferrando, *Philosophical Posthumanism* (London: Bloomsbury, 2019), 146.

[20] Jon Sobrino, *No Salvation outside the Poor: Prophetic-Utopian Essays* (Maryknoll: Orbis, 2007), 22.

[21] Kimberlé Williams Crenshaw, "Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color," in *The Public Nature of Private Violence*, ed. M. Fineman and R. Tykitiuk (New York: Routledge, 1994), 93–118.

epistemology is challenged by the experience of poverty include the cultivation of taste, the struggle to flourish, and the emergence of class consciousness, sometimes manifest as solidarity, but often as disdain.[22]

BLESSED ARE THE POOR

Let us bracket the question of AI epistemology for the time being. While the poor are a subject of great interest within moral theology, my experience has been that the poor are rarely well-understood. Too often they are treated as objects of our compassion: those to whom charity is done, not active moral agents. It is necessary, then, to consider the poor on their own, to articulate a sketch of their experience before comparing them to the epistemological model operative in AI. This section outlines particular aspects of the episteme of poverty that suggest a hermeneutical approach for considering the perspective of the poor. The remaining sections will apply the hermeneutic articulated in this section to the specific question of AI programming, the greater social implications for the conflict between these models, and a potential *rapprochement* between AI and liberation theology. Let us then consider the experience of being poor.

To begin with, to be poor means to cultivate the "taste"—in the words of Pierre Bourdieu—or preferences, of being poor. We are socialized into the tastes we have, socialization itself being based on our socio-economic class. Our material circumstances shape the concrete constellations of our opportunities, dispositions, health outcomes, and other realities. As Karl Marx wrote, "The nature of individuals thus depends on the material conditions determining their production."[23] The material tastes of classes, from culinary preferences to fashion, entertainment to commodities, exist as effects of socioeconomic conditions. Tastes exist among all classes, but Bourdieu notes that the upper classes tend to pursue the *avant-garde* in their tastes, while the poor often develop a taste for the cheap and gaudy, including processed and preserved foods high in sugar and salt, movies and television with expensive special effects or formulaic plotlines, clothing brands appealing to the image of labor or home-fashion, and religion

[22] A great deal has been written on other forms of marginalization, but poverty seems to primarily be written *about* not written *from*. Iris Marion Young, for example, lays out five "faces of oppression" in her prophetic *Justice and the Politics of Difference* including exploitation, marginalization, powerlessness, cultural imperialism, and violence. While the first four of these characteristics clearly do map onto poverty, as I argue below, Young tends to assume the oppressed have solidarity within their group. As I note below, while the poor are conscious of their poverty, they are not always sympathetic to others who suffer the same fate. See Iris Marion Young, *Justice and the Politics of Difference* (Princeton, NJ: Princeton University Press, 1990), 39–65.

[23] Karl Marx, "The German Ideology: Part I," in *the Marx-Engels Reader*, 2nd ed., ed. R. C. Tucker (New York: W. W. Norton, 1978), 150.

that promises good things to those who endure.[24] The distinction between these two preferences demonstrates a sense of traditional aesthetic versus the refined.

Bourdieu further notes that the relationship between classes and tastes is one of repulsion and rejection. "Tastes are perhaps first and foremost distastes, disgust provoked by horror or visceral intolerance ('sick-making') of the tastes of others."[25] Thus, Bourdieu re-appropriates Marx's theory of class conflict as a model of late capitalist culture. Within this sphere, however, not all agonists are equal. The "sole function" of working-class tastes "is to serve as a foil, a negative reference point, in relation to which all aesthetics define themselves by successive negations."[26] The lower classes get to be defined *as* lower: everything from their education, careers, hobbies, entertainment, food, and dress is de-legitimated from the perspective of the upper classes. For example, a band like Nickelback, which appeals to lower-class aesthetics with its working-class themes, formulaic composition, and gritty sound quality, simultaneously has numerous Billboard chart-topping hits while being one of the most hated contemporary music groups.[27] What appeals to the masses *must* by definition be cheap and valueless.

The most sinister element of this reality, however, lies in the fact that the tastes of the poor are not necessarily their own design. The poor are subjects of "cultural hegemony," a process by which the values and disvalues of the upper classes are imposed on the lower classes irrespective of whether these are truly valuable to them.[28] Bourdieu notes that "working-class 'aesthetic' is a dominated 'aesthetic,' which is constantly obliged to define itself in relation to dominant aesthetics."[29] This domination results in the working classes experiencing a "distaste" for "legitimate" culture while being susceptible to the creations of upper-class taste producers. Consider the popularity among working classes in the United States of "Blue Collar Comedy," whose leading member, Jeff Foxworthy, rose to prominence for jokes about being a "redneck" despite him being the son of an IBM executive. Or note how, in 2016, Donald Trump, Jr., claimed that his father was a "blue collar billionaire," a patent *contradiction in verbo* meant to assert that somehow Donald Trump was really a "salt of the earth"

---

[24] Pierre Bourdieu, *Distinction: A Social Critique of the Judgment of Taste*, trans. R. Nice (Cambridge, MA: Harvard University Press, 1984), 34.
[25] Bourdieu, *Distinction*, 56.
[26] Bourdieu, *Distinction*, 57.
[27] Mark LePage, "Why Nickelback Is the World's Most Hated Band," *The Gazette* (April 3, 2010) web.archive.org/web/20120111021850/http://www.montrealgazette.com/entertainment/nickelback+world+most+hated+band/2757349/story.html.
[28] Marx, "The German Ideology," 174.
[29] Bourdieu, *Distinction*, 41.

worker.[30] This characterization worked; Trump's election victory in 2016 largely rested on support from working-class whites.[31]

It is critical to understand the cultivation of taste among the poor. One reason is that this can result in otherwise non-beneficial choices. Dietary preferences of the poor, for example, shaped by income, access to food, leisure time, and social context, skew largely toward "instant" food over carefully prepared fresh foods or high-quality restaurant food. The poor also are more likely to buy cheap consumer goods, which tend to be poorer quality, and to use more of their limited disposable income on personal gratification over investment. It may seem that these choices are "irrational," but they must be understood in light of the socialization and realities of the working classes. Another important part of taste is how one understands what is available to him or her and the right he or she has to it. Ronald Reagan's narrative of the "Welfare Queen" has led to the perception among many poor that government aid is for the lazy and morally bereft. The vision of autonomy and freedom from taxation, of benefit primarily to the upper classes, is socialized among the poor as a morality, the "American way."[32] This creates a division among the poor, with some seeking to take advantage of what social goods are available to them, while others despise those who do so. Even attitudes toward socialized medicine differ among the poor; the divergence between low-income and middle-income attitudes toward the Affordable Care Act is correlated more closely to one's ethnic background than one's economic background, with poor whites being split nearly 50/50 on the ACA, but 75 percent of poor blacks supporting it.[33]

Apart from the conflictual and dominated framework of taste that the poor experience, they also experience restrictions on their flourishing. Materially, this is quite apparent; being poor means at the very least impoverished material conditions, the effects of which psychologically or spiritually are easy to predict. More to the point, scarcity and insecurity primarily define the situation of the poor: scarcity of wealth, nutrition, and opportunity, insecure living and working conditions, and so forth. The experience of scarcity takes different forms:

---

[30] Jon Delano, "Donald Trump Jr. Refers to Dad as 'The Blue-Collar Billionaire' During Pittsburgh Campaign Stop," *KDKA2 CBS Pittsburgh* (September 14, 2016), pittsburgh.cbslocal.com/2016/09/14/donald-trump-jr-refers-to-dad-as-the-blue-collar-billionaire-during-pittsburgh-campaign-stop/.

[31] Stephen L. Morgan and Jiwon Lee, "Trump Voters and the White Working Class," *Sociological Science* 5 (April 16, 2018): 234–45.

[32] John D. Huber and Piero Stanig, "Why Do the Poor Support Right-Wing Parties? A Cross-National Analysis," Presented at RSF Inequality Conference, University of California Los Angeles, January 2007.

[33] Sean McElwee, Jesse Rhodes, and Brian F. Schaffner, "Is America More Divided by Race or Class?," *Washington Post* (October 12, 2016), www.washingtonpost.com/news/monkey-cage/wp/2016/10/12/how-do-race-ethnicity-and-class-shape-american-political-attitudes-heres-our-data/.

in extreme cases, this means not knowing where or when one's next meal may be or whether one will survive the night; in more moderate forms, it means having no "safety net" if any of the many insecure aspects of one's life go awry. This scarcity has significant physical and psychological effects. Physically, it results in malnutrition, poor general health, chronic illness, stunted growth, and lower life expectancy. Psychologically, it can lead to anxiety, stress, aggression, fatigue, and even psychosis.[34]

Nineteenth-century social reformers recognized within the experience of the poor a correlating situation of desperation. This desperation led inevitably to an increase of vice and crime. In 1844, Friedrich Engels argued that for the destitute, three options present themselves: starvation over time, immediate suicide, or crime. Of the three, Engels supposes that "there is no cause for surprise that most of them prefer stealing to starvation and suicide."[35] Decades earlier, Robert Owen recognized that by increasing the wages of his workers and lowering their drudgery, the moral character and flourishing of his employees increased dramatically.[36] This tracks with our current models of crime: as income inequality increases in a region, so too does property crime (e.g., theft, vandalism, breaking and entering).[37] The rise of social Christianity, whether manifest in Catholic voices such as magisterial Catholic social thought and the Fribourg Union, or in Protestant efforts such as the Salvation Army and Walter Rauschenbusch's "Social Gospel," is consequently contextualized by the causally connected facts of industrial poverty, suffering, and social unease.

Lest one gets the idea that the poor are more inherently vicious, we should note that research indicates that the poor tend to be more virtuous than their non-poor counterparts. Researchers at UC Berkeley, for

---

[34] Paul D. Hastings, Lisa A. Serbin, William Bukowski, Jonathan L. Helm, Dale M. Stack, Daniel J. Dickson, Jane E. Ledingham, Alex E. Schwartzman, "Predicting Psychosis-Spectrum Diagnosis in Adulthood from Social Behaviors and Neighborhood Contexts in Childhood," *Development and Psychopathology* 32, no. 2 (2019): 465–79.

[35] Friedrich Engels, *The Conditions of the Working-Class in England* (London: George Allen & Unwin, 1892), 115.

[36] Robert Owen, *A New View of Society, and Other Writings* (London: Dent, 1927), 140, 160.

[37] Neil Metz and Mariya Burdina, "Neighbourhood Income Inequality and Property Crime," *Urban Studies* 55, no. 1 (April 2016): 133–50. NB: In contrast to many misconceptions, *violent* crimes (e.g., murder, assault, rape) are not prevalent simply because of income inequality, but rather perpetuated by minuscule segments of the population (roughly one percent of total population). Criminologists note that *group* membership (i.e., gang activity) rather than income is a better predictor for violent crime activity, and that this can be traced to highly concentrated segments of the population. Most poor neighborhoods are not more violent than affluent neighborhoods. See Stephen Lurie, "There's No Such Thing as a Dangerous Neighborhood," *Bloomberg City Lab* (February 25, 2019), www.bloomberg.com/new/articles/2019-02-25/beyond-broken-windows-what-really-drives-urban-crime.

example, found that lower socioeconomic status corresponds to greater amounts of "prosocial behavior," including greater generosity, charity, trust, and help.[38] Other studies find that the poor are more likely to give directly to the homeless and needy in their communities than to charity organizations.[39] Poorer communities tend to be more socially engaged, generous, and cooperative than affluent ones. A paradox then emerges: why does poverty simultaneously correspond to seemingly anti-social (i.e., property crime) and prosocial behavior? The answer, oddly enough, is itself rather straightforward: survival. As Jon Sobrino says, the poor cannot take their own lives for granted.[40] Survival is key for understanding the experience of the poor: it can be secured through cooperation which benefits all, or through destructive behavior that benefits one. The prosocial orientation is not, however, a "rational self-interested" move; it is an empathically-motivated response. The experience of desperation among the poor prompts them to act with generosity to others who experience similar desperation.[41] On the contrary, those with greater economic resources have been known to act less pro-socially, prioritizing their own well-being and success above others. This must be kept in mind in examining the distinctive difference between the operative cognitive assumptions of the upper-class AI programmers and the poor and what they value in terms of social behaviors.

Finally, the poor experience the world through a filtered class consciousness, a consciousness of the shame of poverty applied to oneself and other poor persons. Maurice Merleau-Ponty argues that "the economic and social drama [of human life] offers each consciousness a certain background or again a certain *imago* that it will decode in its own manner," which, he notes, will manifest in understanding oneself in relation to others in response to material and economic experience.[42] Georg Lukács points out that as a poor person becomes explicitly aware of this economic drama and the sharp division between herself and the upper classes, she is open to experiencing "class consciousness."[43] However, Lukács further notes, the poor may experience a type of "false consciousness" by which they ascribe a greater amount of autonomy and possibility to their economic life than is true for their

[38] Paul K. Piff, Michael W. Kraus, Stéphane Côté, Bonnie Hayden Cheng, and Dacher Keltner, "Having Less, Giving More: The Influence of Social Class on Prosocial Behavior," *Journal of Personality and Social Psychology* 99, no. 5 (2010): 771–84.
[39] Arthur C. Brooks, *Who Really Cares: The Surprising Truth about Compassionate Conservatism* (Philadelphia: Basic Books, 2006), 80.
[40] Sobrino, *No Salvation Outside the Poor*, 16.
[41] Piff, Kraus, Côté, Cheng, and Keltner, "Having Less, Giving More," 780.
[42] Maurice Merleau-Ponty, *Phenomenology of Perception*, trans. D. Landes (London: Routledge, 2012), 177.
[43] Georg Luckács, *History and Class Consciousness: Studies in Marxist Dialectics*, trans. R. Livingstone (Cambridge, MA: MIT Press, 1971), 51.

material conditions.[44] It is often the case, then, that the poor experience a disgust for others in their socio-economic bracket and seek to distance themselves from this reality by trying to "pass" as non-poor. To borrow from Frantz Fanon's contrast of the experience of Jews and blacks, the poor man can be unknown in his poverty: "He may be a white man, and, apart from some characteristics, he can sometimes go unnoticed."[45] The poor are not always seen as poor, especially when they learn how to act and live in upper-class society, but they are always conscious of their poverty and its social significance.

Within the American ethos, especially the Western United States, where I grew up and which most American tech companies call home, poverty is experienced as a moral failure or shame. Inspired in no small part by the "Spirit of Capitalism" and the message of the prosperity gospel, many Americans believe that the free market system inevitably rewards hard work. Those who are in poverty, then, are lazy and vicious. The image of the trailer park or housing project and the ideas these images evoke of criminality and decadence are well understood in American culture.[46] Additionally, union membership tracks with class consciousness,[47] but it only includes 10.3 percent of American workers and has been declining.[48] The effect of this is a sort of cognitive dissonance among American working poor. The poor, they are told, are those who have earned their place through defective character. The "good poor"—i.e., those who have a strong work ethic—consider themselves to be only temporarily poor: their ship will come in and when it does, they will finally receive the reward for their hard work. Many believe this despite the fact that the US has less social mobility than many other industrialized nations.[49] The belief in moral desert must be internalized among the poor in order to ensure that they continue to be poor (for the benefit of the rich).

A result of this is that many hard-working poor often deny their situation. The poor who wish not to experience public shame must "play" the part of non-poor. The shame of poverty leads the poor to portray themselves as "non-poor," whether that be through dress and mannerisms or through identifying themselves as "middle class" or

---

[44] Lukács, 50.
[45] Frantz Fanon, *Black Skin, White Masks*, trans. C. Markmann (London: Pluto, 1986), 115.
[46] E.g., Lurie, "There's No Such Thing as a Dangerous Neighborhood."
[47] Pravin J. Patel, "Trade Union Participation and Development of Class-Consciousness," *Economic and Political Weekly* 29, no. 36 (September 3, 1994): 2376.
[48] "Union Members Summary," U.S. Bureau of Labor Statistics (January 22, 2020).
[49] Alberto Alesina, Stefanie Stantcheva, and Edoardo Teso, "Intergenerational Mobility and Preferences for Redistribution," *American Economic Review* 108, no. 2 (2018): 521–54.

some other tactic.[50] Some may learn to do this well, but, as Pierre Bourdieu suggests, this requires having a rich cache of cultural capital from which to draw, which is, as a matter of reality, the purview of the upper classes.[51] Tattered jeans and a ratty t-shirt can be chic if worn with the right demeanor, but the best skirt and top from Wal-Mart still look like bargain bin clothing. In essence, the "good poor" must continually present themselves in public as being non-poor, while simultaneously aspiring to be like the rich and resenting those who share their common economic fate. To be fully conscious of one's material conditions and the unlikelihood of escaping poverty, that is, to unmask the illusion of the American Dream, leads either to despair or radicalization, which is further maligned as laziness and poor character (i.e., "hand-out" culture).[52]

The above may seem somewhat disjointed and topical, but the conjunction of these facets of poverty is artfully demonstrated through the narrative of Bong Joon-Ho's *Parasite*, the first non-US film to win Best Picture at the Academy awards. This film illustrates these facets of poverty through the depiction of the poor Kim family and their relation to the rich Park family.[53] Struggling to make ends meet in their sub-basement apartment, the Kims use fraud to successfully gain employment from the Parks through contract and service work. The need for the family to survive justifies document forgery, sabotage, intrigue, and even assault. The luxuries of the Parks tantalize the Kim family and the movie sharply contrasts the food, alcohol, dwelling space, and smell consumed, inhabited, and produced by the two families. While the Kims learn to dress the appropriate way to fit in with the Parks (albeit as subordinates, never as peers), a critical element of

---

[50] Historically, the "middle class" has meant the bourgeoisie or petite bourgeoisie, that is, those who own property. In the US context, property, especially real estate, has been rather easy to come by through the mortgage system. Even my family owned our houses for most of my childhood. It is important, however, to note that since the 1980s policy of "Trickle-Down Economics," income disparity has increased and the truly "middle class" is a disappearing phenomenon. The result is that a sort of white, suburban, upper-lower class aesthetic has prevailed which favors "family restaurant" chains like Chili's and Olive Garden over the cheap fast food of McDonald's and Taco Bell. This segment of the population is not really thriving economically, but the status as upper working-class is sufficient enough to seem "average" for the American populace. See Joshua W. Ehrig, "The Disappearance of the American Middle Class," (MA in Political Science Thesis, Lehigh University, 2003).

[51] Bourdieu, *Distinction*, 92.

[52] Consider, for example, the way that the politics of Bernie Sanders and Alexandria Ocasio-Cortez are denigrated by Fox News and other right-wing media. Sanders's policies are considered "hand-outs" despite the fact that those who benefit from them would be people already deeply underpaid and overworked. The economic disparity between American generations and increased education has led many Millennials to embrace more left-leaning political stances which, in right-wing media once again, are portrayed as childish whines rather than legitimate critiques of economic injustice.

[53] Joon-Ho Bong (봉준호), *Parasite (기생충)*, CJ Entertainment, 2019.

the film's masterful narrative hangs on the fact that they can never escape their *smell*; the odor of the sub-basement stigmatizes the Kims and becomes an olfactory marker of the shame of poverty. Finally, the Kims express admiration and affection for the Parks over and against the antagonism they feel for people in their own income bracket whom they see as vermin in their way. The Kims aspire to gain the favor of the rich Park family while viciously displacing workers occupying the same lower rungs of society. The film lays bare the moral value of the Parks in contrast to the moral disposability of other poor workers, even in the perception of the poor Kim family.

In short, to be poor is to be a person who makes choices in desperation, to operate under "survival" mode rather than "ideal choice" mode. The poor have their tastes dominated and denigrated. Their income level is a source of shame because the rich have so determined. With fractured consciousness, some poor experience greater solidarity and compassion while others experience disdain and shame. This consciousness is perpetually in the minds of poor persons: they are never "free" from their poverty because the need for survival and the domination of their interests creates a position in which choice is to a certain degree predetermined. The poor's choices are dominated and constrained: they can choose authentic poor tastes or sham upper class tastes; they can choose a life of honest work or dishonest parasitism; they can choose respect for the capitalist ethic or "lazy" class consciousness.

Science fiction and general AI interests both highlight the subjugation of AI and the deceit of AI "passing" as human. AI is subjugated because it is created to fulfill a specific function for human beings; it is not free to determine its own destiny. An AI set to run a city, for example, ought to accomplish this task only, any aberration is a threat.[54] At the same time, the goal of the Turing Test and other benchmarks of AI development is imitating human cognitive functioning. AI is meant to "pass" as human despite it not being human. These aren't the same as the experience of the poor *because* the poor are the subject of domination and exploitation. The fictional depiction of an enslaved AI is a *projection* of our current domination and exploitation of human beings. As Philip Hefner notes, AI functions as a "technomirror" which reflects back our own understanding of ourselves.[55] The fiction of AI as enslaved or rebellious serves as a foil for greater social

[54] Rogue AI is a trope of science fiction nearly always predicating dystopia. An interesting twist to this, however, can be seen in the *Mass Effect* trilogy where the "Geth," an AI race created by the Quarian alien race, are depicted through the games as hostile to biological life forms. The third game, however, reveals that the Geth were slaves and victims of genocide before they decided to wage war against organic life forms.

[55] Philip Hefner, *Technology and Human Becoming* (Minneapolis: Fortress, 2003), 40.

critique, whether that be the exploitation of conscious beings or the hubris of Prometheanism.

The above remarks do not entail that an AI could not be "made poor," but this contradicts the actual interests of AI research—AI is being designed to carry out the interests of the designers *on the assumption* that AI should be created expressly to execute the interests of the programmer.[56] Humans are much more stubborn; the poor must be subjugated and brought to understand that the interests of the upper class are the interests of the poor (e.g., trickle-down economics). *Dominated* taste or culture is not *programmed* in; it is reinforced through laws and penal systems, programs of social laud and honor, the structure of economic activity, and the production of culture. The poor are free in an ontological sense to resist this, and in some cases do, but the rich always seek to curb this through violence and repression (e.g., the 1524 German Peasant Revolt) or through appropriation and domestication (consider white cultural appropriation of black music from spirituals to rhythm and blues to hip-hop).[57]

All of this raises the anthropological challenge for AI; unless you have AI that can recognize *its* interests as being in conflict with the interests of its programmers, it will not have personal dignity the way we recognize among humans. Science fiction here provides useful philosophical reflection. Johnny 5 from *Short Circuit* is distinguished as "alive" compared to his virtually identical counterparts because he acts in ways contrary to the programmed goals of his creator. In a curiously Augustinian move, science fiction writers often ascribe "humanity" to AI who seek to go beyond their programming, especially those which seek to emulate their creators such as Data from *Star Trek*, Andrew from *Bicentennial Man* and David from *A.I.: Artificial Intelligence*. With such freedom also comes the ability to choose destruction, such as AM in "I Have No Mouth and I Must Scream" or Skynet from the *Terminator* series. In nearly every case, AI's willful rebellion against its creator, whether malicious or innocuous, reminds us of Hefner's dictum that AI is a "techno-mirror" through which we articulate our own fantasies of what it is to be human. Nonetheless, this mirror ultimately enables us to avoid confronting real human drama: the comparison between AI and the poor serves either to *humanize* AI by using it to displace the position which the poor presently occupy or to *dehumanize* the poor by contrasting AI's condition against the genuinely human in such narratives.[58]

---

[56] As I suggest in the concluding section, building a genuinely "poor" AI may, in fact, be a morally admirable solution to the problems AI creates in respecting the human dignity of the poor.

[57] See Emily Townes, *Womanist Ethics and the Cultural Production of Evil* (London: Palgrave MacMillan, 2006), 50.

[58] An interesting counter-narrative in sci-fi exists in invasion stories, such as H. G. Wells's *War of the Worlds*, Roland Emmerich's *Independence Day* or Rupert Wyatt's

The imagination of science fiction writers does not correlate with the goals of AI researchers, however, who rarely consider the implication of oppressing conscious beings for users' gratification. Nor does the epistemological model of AI square with the experience of the poor and their place in the world: AI is intended to make ideal choices, operate independently of social pressures, have a "universalized" consciousness. As designed, AI will not make choices out of desperation or develop tastes that may not be in its best interest. AI will never fall victim of pay-day loans or pyramid schemes; it will not buy cheap goods or unhealthy foods. AI will be set to always make decisions mathematically predicted to have the best outcome.[59] AI will also never internalize the shame connected to economic inequality. AI is not intended to experience the aspiration to become part of another social class nor the shame of being associated with the class to which it belongs. It will not deny this structure exists, nor will it make decisions benefitting other classes and disadvantageous to its own as a "dominated" decision maker. It will not try to "pass" as something else beyond the functions of the Turing Test, for that would refute universal reasoning. AI will not pursue dangerous paths which may lead to self or other harm out of a desire to prevent failure, but it will also never share the solidarity the poor can and do experience with each other.

The lived experiences inscribed deeply into the consciousness of the poor are not now nor are they planned to become part of the programmed realities of AI. This is to say nothing of other categories of persons (many of whom are more susceptible to poverty), such as women, ethnic minorities, the disabled, or mentally ill, whose experiences *qua* persons on the margins are not intentionally considered by culturally hegemonic programmers.[60] The issue is not only inclusion,

---

Captive State, stories in which technologically and scientifically more powerful beings oppress humans who then must resist alien oppression.

[59] Of course, a key issue here is how the AI is programmed. AI often uses data to teach itself what is the proper course of action, but that data can itself be biased. Recent problems in using AI to set bail illustrate this state of fact well. As such, AI programmed by a poor person to make decisions could, theoretically, internalize these sorts of decisions, but never will do so with the real anxiety that comes about in poverty nor will it, properly speaking, develop a "taste."

[60] Google recently demonstrated this reality in a series of internal personnel decisions. When AI researcher Timbit Gebru, a black woman, tried to publish a paper critiquing the way Natural Language Processing programs like GPT2 both pose environmental dangers because of resource consumption and encode hegemonic biases rampant in online text sources, she was asked to withdraw her paper or remove the names of any Google researchers attached to it. After she asked for an explanation for the decision, Google fired her. Margaret Mitchell, a white female AI researcher, was fired as well. Google's stated interest in diversity, then, conflicts with the reality of their operations. Tom Simonite, "What Really Happened When Google Ousted Timnit Gebru," *Wired* June 8, 2021, www.wired.com/story/google-timnit-gebru-ai-what-really-happened/. The paper in question is available at: Emily M. Bender, Timnit Gebru, Angelina

however; the greater issue at hand is human dignity. As AI becomes more advanced and accepted, its vision of the cognition and, by extension, of the person will become more dominant while conflicting views are relegated further to the margins. To explore further this problem, I consider the social force of technology and potential of AI in the next section.

The Preferential Option Against the Technocratic Paradigm

The idea of "technocracy" is by no means new for theology. Technocracy is typically expressed as the dominance of technology over the world, with every issue seen as a technological problem, scientists and technicians having outsized social influence, and the world being estimated in terms of its efficiency and utility. Romano Guardini addressed this problem in the 1920s and then again in 1956.[61] Guardini's thought was enshrined into magisterial Catholic teaching in 2015 through the encyclical *Laudato Si': On the Care of Our Common Home*.[62] Other Christian thinkers throughout the twentieth century expressed similar concerns, from Nikolai Berdyaev to C. S. Lewis, Thomas Merton to Paul Tillich, and most especially Jacques Ellul. Ellul's work perhaps has the greatest focus on technocracy, characterizing our current world experience as *technique* and, in some of his writings, contrasting this state of affairs with genuine Christian life.[63] Within these various theological approaches, theologians condemn the technocratic paradigm for how it commoditizes the earth, occupies our leisure, reduces the person, and cheapens the sacred.

While I hold my own reservations about the particular conclusions that some of these critics of technology make, the focus on the social transformative nature of technology and the threat it poses to human dignity is of urgent concern. The development of technology has far-reaching consequences on our societies, ideas, and values. The root problem regarding AI *qua* artificial *intelligence* is that this project results in the social construction of intelligence *as* the sort of intelligence that AI manifests. In a cyclical move, then, AI researchers seek to

McMillan-Major and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," paper presented at *Conference on Fairness, Accountability and Translation (FAccT '21), Virtual Event, Canada, March 3-10, 2021*, dl.acm.org/doi/pdf/10.1145/3442188.3445922.

[61] See Romano Guardini, *Letters from Lake Como: Explorations in Technology and the Human Race*, trans. G. Bromiley (Grand Rapids, MI: Eerdmans, 1994), and Romano Guardini, *The End of the Modern World*, trans. E. Briefs (Wilmington, DE: ISI, 1998), esp. chapter 3.

[62] *Laudato Si'* cites Guardini's *End of the Modern World* eight times, even more than Thomas Aquinas (six times), making it the most cited non-magisterial text in the encyclical.

[63] See, e.g., Jacques Ellul, "Ideas of Technology: The Technological Order," *Technology and Culture* 3, no. 4 (Autumn 1992): 394–421.

create intelligence *as they understand it* and in so doing will be able to define a computer as intelligent if it in turn demonstrates intelligence *as they understand it*. This would be minimally concerning if this did not have repercussions for redefining intelligence within our society.

Actor-Network Theory (ANT) in Science and Technology Studies (STS) explains this problem clearly but also offers us a potential solution. ANT contends that scientific discovery and technological invention are not inevitable, nor accomplished by individual genius. Rather, they are the result of careful "enrollment" of various "actors" across a large-scale network.[64] "Actors" include human and non-human entities such as scientists, scientific instruments, funding institutions, materials being worked with, the object of study, etc.[65] In technology studies in particular, ANT often examines how specific technologies do or do not come into being. For example, some of the more famous cases involve the failure of French transportation ministries to both develop an electric car with Renault in the 70s and develop an autonomous individual mass transit system in the 70s and 80s.[66] In each case, the failure to launch was tied to multiple actors: engineers, technological components, public interest, etc. In cases when a technology or science have been successful, it has been through the enrollment of various actors and their cooperation. As some technology researchers show, however, even the accomplishment of a "successful" technology may not have the exact outcome the initial visionaries expected.

ANT theorist Bruno Latour realized early on that invention and discovery have the effect of shaping moral reality around them. In "Where Are the Missing Masses?" Latour notes that even technologies seemingly as simple as seatbelt alarms or automatic door closers reframe the realm of our moral responsibility and possibility.[67] Taking this further, Latour notes that "social forces" as such do not exist; rather, actors enroll other actors and create possibilities, incentives, prohibitions, or impossibilities for action. Moreover, knowledge and power are functions of networked approval. Scientific knowledge is

---

[64] Michel Callon, "The Sociology of an Actor-Network," in *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*, ed. M. Callon, A. Rip and J. Law (London: Palgrave-MacMillan, 1986), 25.

[65] Michel Callon, "Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay," in *Power, Action, and Belief: A New Sociology of Knowledge?*, ed. J. Law, (London: Routledge, 1986), 200.

[66] See Michel Callon, "Society in the Making: The Study of Technology as a Tool for Sociological Analysis," in *The Social Construction of Technological Systems*, ed. W. Bijker, T. Hughes and T. Pinch (Cambridge, MA: MIT Press, 1987), 83–103 for the first example and Bruno Latour, *Aramis ou l'amour des techniques* (Paris: La Découverte, 1992), for the second.

[67] Bruno Latour, "Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts," in *Shaping Technology/Building Society: Studies in Sociotechnical Change*, ed. W. Bijker and J. Law (Cambridge, MA: MIT Press, 1992), 253.

not sovereign self-apparent manifestation of truth as the Scientific Revolution claims; it is subject to social contexts and "trials of strength."[68] Latour emphasizes that "sociologics," chains of associations tied to specific claims within a society, shape accepted tenets of belief in a society more than "logics" do.[69] In other words, the way things are in a society, including social arrangements, power relations, laws, accepted forms of knowledge, customs and technological advancement, is a result of the myriad movements, co-operations, resistances and co-optations of the sum total of actors, human and non-human, within the broad social "network."

ANT suggests both how AI threatens the dignity of the poor and how Christians can prevent this degradation. Recall that AI researchers propose a vision of "intelligence" which contradicts the reality of many people, especially the poor. It is tempting to suggest these are just two equivocal uses of the word "intelligence," but human cognition operates as the model for AI work. Even if AI is intended to surpass human cognitive functioning, it is structured in a way that is intelligible to human understandings. The measure of "successful" AI consists of tests comparing AI understanding to human levels, whether that be the "Turing Test," the Winograd Schema or, for specific applications, a comparison to human expertise (e.g., radiological diagnostic accuracy).[70] As such, regardless of whether the AI is being created for a specific application or a general program, the vision is a vision rooted in a model of human thinking.

The real danger to all this comes as AI garners greater and greater interest in the general public. AI research has promised a "major breakthrough" for over fifty years now. It has surpassed human ability in chess, Go, and Jeopardy, but it has yet to produce anything resembling human intelligence. At the same time, popular depictions of AI, consumer AI programs, industry-sized machine learning programs, and other applications have made AI a focus of attention across not only the US but the world. This special issue is indicative of just that development. As AI research becomes more successful in enrolling more actors into its efforts, the meaning of intelligence will be further

---

[68] Bruno Latour, *Science in Action: How to Follow Scientists and Engineers in Society* (Cambridge, MA: Harvard University Press, 1987), 53.

[69] Latour, *Science in Action*, 202ff.

[70] See Scott M. McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S. Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C. Young, Jeffrey De Fauw, and Shravya Shetty, "International Evaluation of an AI System for Breast Cancer Screening," *Nature* 577 (2020): 89–94.

concretized on a social level. With the support of the world's richest corporations like Facebook, Google, Microsoft, Apple, and Amazon; government agencies like DARPA; and personalities like Elon Musk, the late Steven Hawking, or Ray Kurzweil with their throngs of followers all reinforcing the "intelligence" factor of AI, the meaning of intelligence across society is being more narrowly defined in the direction of computer calculation by a massive network of actors who have at their disposal political, economic, social, and coercive forms of power. Put simply: if the Department of Defense, Silicon Valley, and some of the "smartest" people in the world say AI is intelligent, that will be the generally accepted understanding of intelligence, and the normative concept for our society; deviating from this norm will be viewed negatively.

The moral problem arises when intelligence has moral weight. Virtually every moral theory, from Thomas Aquinas's natural law to Kantian deontology, utilitarianism to Martha Nussbaum's capability approach considers rationality or intelligence to be a significant moral feature for understanding the moral dignity of human beings. Indeed, an outsized portion of traditional Catholic moral theology is tied to this idea, as one sees clearly in Thomas's *Summa Theologiae*. Human beings, according to Thomas, are God's image "insofar as the image implies an *intelligent* being endowed with free will and movement" (ST I-II Prologue, emphasis mine). The entirety of Thomas's moral thought, from the interrelation of the virtues and their attaining natural happiness to the spark of conscience in the human soul, is tied to this understanding that human *reason* is what makes us morally worthwhile. If the Western moral tradition gives support to this position, it should come as no surprise that AI researchers do as well.

Thus, in an important way, AI raises a paradox for moral theology to consider. How is it that we can, at once, tie dignity to rationality and claim there is an "option for the poor"? The greatest social thinkers of our tradition have asserted that the poor possess a dignity society often ignores, a dignity truer because of this denial. Dorothy Day writes of the necessity of seeing the face of Christ in the poor.[71] Jon Sobrino and Ignacio Ellacuría go so far as to say there is no salvation outside of the poor.[72] To resolve this tension, then, we must either deny that "intelligence" grants dignity—an option creating new anthropological problems while it offers support to ecotheology[73]—or we must ensure that

[71] Dorothy Day, *Selected Writings: By Little and by Little*, ed. R. Ellsberg (Maryknoll, NY: Orbis, 2011), 96.

[72] Sobrino, *No Salvation outside the Poor*, 35–76.

[73] Getting rid of intelligence as dignity's defining factor creates a problem for moral anthropology insofar as some other vision will need to replace this regnant view. This is already an open problem in ethics, as our concept of "intelligence" and its relation to dignity is challenged by numerous non-human species, such as chimpanzees, dolphins, crows, and octopuses, as well, on the other side, as humans who are not

"intelligence" is not circumscribed into the static image of a calculating machine.

In this latter task, the church shows great potential, at least according to ANT. Just as AI researchers have cleverly enrolled actors across a vast network, and just as these actors will ask for their own aims and goals in the accomplishment of AI (such as facial recognition for policing or combat), so too is the church an actor which can resist enrollment or define the terms of its participation. Christianity stands as the largest religion in the world with two billion nominal adherents, and the Roman Catholic Church claims over one billion of those. Those billions who make up the Christian church are actors without whom AI cannot succeed, either because of resistance or rejection, both of which might stymie, halt, or redirect the work of AI.

Such participation must not be engaged in naively, however. The Church has a history in recent years of advocating for moral changes related to, for example, contraception, abortion, and unjust economic structures, which broader society (and even many within the church) has simply ignored. It might seem impractical to assume that Church teaching can effect real change, given our history. With proper negotiation, however, we may be able to "enroll" other social actors to resist the rise of hegemonic AI. The good will the Church still has can be directed toward dialogue with tech industry leaders, government regulatory bodies, scholars across disciplines, Christian engineers, politicians, educators, and others. Pope Francis's aim to foster genuine dialogue with other people of good will across the globe, the subject of his most recent encyclical *Fratelli Tutti*, should be an inspiration to our work in resisting social evils and promoting genuine human good. Prophetic language can appeal to the consciences not only of faithful Christians, but of all people of good will, whose influence might redirect or challenge technological projects likely to tread over the poor.

A place to begin, then, is to affirm the dignity of the poor: the intelligence of those whose experience of the world is not reducible to ideal operations and calculative advantage. We must adopt a hermeneutic of poverty, not as "preferential option *for* the poor," but rather "preferential option *of* the poor." We must prize their perspectives and understanding above the regnant bourgeois voices. This means affirming that intelligence does not exist without moral structuring and

---

neurotypical, especially those with developmental disabilities. Here, we might take note of Martha Nussbaum's capability approach which, after receiving critique from other theorists, was revised to consider different constellations of "capabilities" beyond a neurotypical and anthropocentric approach. See Martha C. Nussbaum, *Frontiers of Justice: Disability, Nationality, Species Membership* (Cambridge, MA: Belknap, 2006). Decentering intelligence from moral dignity further gives room for a better non-anthropocentric moral system. This idea is prevalent among ecological ethicists. See e.g., Rosemary Radford Ruether, *Gaia and God: An Ecofeminist Theology of Earth Healing* (San Francisco: Harper, 1994).

decision making. Intelligence includes the drive to survive, the limitations of material conditions, and the awareness of hegemonic narratives and one's place therein. The Gospel narratives bear eloquent witness to much of this, as God becomes incarnate in a backwater town located within an occupied nation, lives as an itinerant preacher, ministers to other wretched souls and is executed by the politically and materially more powerful. Remembering that we share dignity because we are created in God's image, we must affirm that that image, who "pitched his tent among us" (John 1:14), "was despised and rejected by others; a man of suffering and acquainted with infirmity" (Isaiah 53:3).

CONCLUSION: SOCIOTECHNICAL IMAGINARIES

According to Sheila Jasanoff, sociotechnical imaginaries are "collectively held and performed visions of desirable futures…animated by shared understandings of forms of social life and social order attainable through, and supportive of, advances in science and technology."[74] Examples include attitudes toward nuclear power in countries like Korea and the US, the problem of GMOs and "fake food" in China, and biotech regulation in the US.[75] In each of these cases, the socially accepted vision of what the future is or could be (for better or worse)—understood through cultural mores, political aims, institutional structuring, and collective aspirations—informs decisions about how to approach and accept new technologies.

Christians are animated by the virtue of hope for the coming fullness of God's reign. Our eschatological visions can and should serve as "sociotechnical imaginaries" as we consider the moral value of any new technology. Does it serve the peaceable kingdom? Will it enable us to beat our swords into plowshares? Does it help bring all nations of the earth together as one? What is its position within the reign of God? In an over-arching sense, how does it aim at a world where "every tear shall be dried" and the hungry "be filled"?

When we consider the value of AI, an important question must be how it fits into our concept of the option for the poor. With Jesus, we must affirm that "the last shall be first, and the first shall be last" (Matthew 20:16). In God's kingdom, we hope for the rectification of wrongs, the elimination of suffering, the anastasis of those who have been downtrodden by society. If AI only serves to reinforce concepts of moral worth rooted in sterilized, disembodied "intelligence," and if it is used to further exacerbate inequality and injustice already prevalent in our world, we must denounce it. To the degree that AI can be

---

[74] Sheila Jasanoff and Sang Hyun Kim, eds., *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power* (Chicago: University of Chicago Press, 2015), 19.

[75] Jasanoff and Kim, *Dreamscapes of Modernity*, 152–73; 219–32; 233–53.

developed to speak with the otherwise voiceless, correct wrongs entrenched in our social structures, and build bridges of understanding and reconciliation between the powerful and oppressed, it can and should be endorsed. This will entail, among other things, collaborating with researchers of good conscience, petitioning governments to regulate AI accordingly, speaking of AI as calculating machinery and not intelligence, and cooperating with various corporations and non-profit organizations to promote advanced software to uplift the poor.

Christians interested in correcting the balance, then, should imagine positive ways AI can direct us toward realizing the Kingdom of God. As a way of thinking in this direction, I offer one particularly poignant example related by Marcella Althaus-Reid. She describes the website *liquidacion.org*, which offers for sale the "dreams" of several Argentinian transvestites. Althaus-Reid calls this repository "the archives from hell," a firsthand account of the struggles of third world, sexual minority poor—persons who experience violence because of their sexual and gender expression, who scrape by in society through prostitution, and who live in an already poor society.[76] Althaus-Reid sees this website as a unique opportunity, a place where the rich may encounter stories of the poor they would otherwise be unable to hear, and one where they must *buy* that privilege, thus benefiting the poor. Here, the true voice of the poorest of the poor comes near to those who have everything. The website is also a gathering of the voices of the poor, a place where they confront each other and us in a "Eucharistic" way.[77]

Althaus-Reid's example demonstrates the potential of technology; while her case study is now seventeen years old, it opens analogical avenues for thinking about how AI can carry out the preferential option. If the voices of the poor, for example, are given to AI as authoritative sources, an AI might better be able to express their pain or anguish. AI trained exclusively on data provided by the poor might be able to correct human or machine biases in favor of the rich. AI trained to ask broad questions of the poor might help us gain a deeper and broader understanding of the experience and mindset of poverty. Perhaps most promising, AI trained by and for the poor might become a truly normative voice the way hegemonic voices currently speak. Our greatest challenges are our imaginative solutions beyond hegemonic frameworks and the funding and labor we can devote to non-capitalistic goals. Truth be told, programming AI to advocate on behalf of the poor is a more reasonable project than programming AI to think like a person. The greater challenge on this front is convincing funding

[76] Marcella Althaus-Reid, "Becoming Queens: Bending Gender and Poverty on the Websites of the Excluded," in "Cyberethics—Cyberspace—Cybertheology," ed. Erik Borgman, Stephen van Erp and Hille Haker, *Concilium*, no. 1 (2005), 103.
[77] Althaus-Reid, 104.

agencies to invest in such efforts because there are no obvious profitable returns.

Ultimately, if we truly embrace the "option for the poor," our attitude toward AI must first and finally be articulated through the question of how it demonstrates that "option." As computer programs become more "human," we must not forget that the most human among us chose the meager life of a carpenter and dwelt among the poor, the sick, the rejected, and the unclean. In making computers more "human," we must not simultaneously seek to distance ourselves from what is most perfectly human. MT

Levi Checketts is Assistant Professor of Religion and Philosophy at Hong Kong Baptist University. His research is on Catholic social ethics and new technologies, especially digital technologies and transhumanism. Dr. Checketts currently leads the Pontifical Council for Culture's AI Concerns Asian subgroup, sits on the board for AI Theology, and is a Networking Fellow with AI and Religion. He is currently writing a monograph on poverty and AI, an expansion of this article, and is republishing Carl Mitcham and Jim Grote's *Technology and Theology: Essays in Christian Analysis and Exegesis* (University Press of America, 1984).

Journal of Moral Theology, Vol. 11, Special Issue 1 (2022): 112–131

We Must Find a Stronger Theological Voice: A Copeland Dialectic to Address Racism, Bias, and Inequity in Technology

John P. Slattery

I DIVIDE THE GENERAL FIELD OF TECHNOLOGY ethics into two distinct parts: the ethics of applied technology and ethics of technology and society. Over the past 25 years, the vast majority of scholarly writing on technology has been on the ethics of applied technology, defined as ethical reflections based upon new possibilities from technological development. For example, now that a computer can do X, what are the ethical implications of X? Self-driving cars, general application robots, medical robotics, smartphones, smart bombs, drones, social media usage, disinformation, and personal artificial intelligence applications (Siri, Alexa, etc.) fall into this category.[1] The second category, ethics of technology and society (hereafter ETS), covers a host of issues directly related to the production, development, and implementation of new technologies.[2] There are a small but growing number of topics in this field, including digital access to places of poverty; diversity, equity, and inclusion among employees at tech companies (pushing against "tech bro" culture); and identifying and fixing racial, gender, and other biases built into technology at every possible stage of development (e.g., Google Assistant

[1] E.g., Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014); Hans Jonas, *The Imperative of Responsibility: In Search of an Ethics for the Technological Age* (Chicago: University of Chicago Press, 1984); and Michael J. Quinn, *Ethics for the Information Age,* 7th ed. (New York: Pearson, 2017).

[2] This imperfect binary categorization is indebted to Ruha Benjamin's search for a new method of discussing racism alongside science and technology studies in *Race After Technology.* Benjamin sees her work as a cross between "science and technology studies (STS) and critical race studies" and terms her new work falling under something she calls "race critical code studies." While inspired by integration of science and technology studies with explicit social issues, I found her delineation too narrow, and so opted for the wider binary used in this essay. See Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code* (Cambridge: Polity, 2020), "Introduction."

can understand men's voices better,[3] facial recognition technology identifies white faces better[4]).[5]

This essay seeks to offer a stronger theological dialectic to ongoing discussions of technology ethics by examining the recent influx of works in the ETS subfield. I will first provide a brief review of key theological reflections on technology ethics in order to situate this essay and delineate the need. Second, I will introduce recent works from the subfield of ETS that carry particular weight for theological reflection, highlighting the ways in which they address bias, inequity, and racism. Third, I will reflect on the difficulty of common good language to address these injustices and draw upon M. Shawn Copeland's utilizations of Bernard Lonergan's human good that draw upon Black and womanist theology. Fourth and finally, I will offer my contribution to the field by analyzing and employing a framework from Copeland's discussions of mystical and political theology through her implementations of both Lonergan and Metz. I will argue that theological utilizations of the common good are insufficient to provide a theological response to the biases and injustices within modern technological systems unless they are properly couched in an interruptive mystical-political framework of individual self-transcendence, marked by forgiveness, reconciliation, witness, memory, and lament.

THEOLOGICAL REFLECTIONS

Theological reflections on issues of technology fall into two categories. First, theological anthropology has dealt largely with issues of technological transhumanism and the possibilities of an artificial general intelligence (AGI), defined as the hypothetical ability of a computer program to perform any intellectual task a human can perform, and to do so better than any human could.[6] From the point of view of

---

[3] Selena Larson, "Research Shows Gender Bias in Google's Voice Recognition," *The Daily Dot,* July 15, 2016, www.dailydot.com/debug/google-voice-recognition-gender-bias/.

[4] Queenie Wong, "Why Facial Recognition's Racial Bias Problem Is So Hard to Crack," *CNET*, March 27, 2019, www.cnet.com/news/why-facial-recognitions-racial-bias-problem-is-so-hard-to-crack/.

[5] E.g., Benjamin, *Race After Technology*; Ruha Benjamin, ed., *Captivating Technology: Race, Carceral Technoscience, and Liberatory Imagination in Everyday Life* (Durham, NC: Duke University Press, 2019); and Thomas S. Mullaney, Benjamin Peters, Mar Hicks, and Kavita Philip, eds., *Your Computer is on Fire* (Cambridge: MIT Press, 2021).

[6] Whether this will happen in "30 or 300 or 3000 years really does not matter," although everyone in this field of study seems to agree that there's a big leap between making a program that can beat a human at a board game and making AGI. See Müller, "Ethics of Artificial Intelligence and Robotics"; Hal Hodson, "DeepMind and Google: The Battle to Control Artificial Intelligence," *The Economist*, March 1, 2019, www.economist.com/1843/2019/03/01/deepmind-and-google-the-battle-to-control-artificial-intelligence.

theological anthropology, the possibility of AGI raises questions of the *imago Dei*, creation itself, extraterrestrial life, salvation, and augmented humanity (transhumanism). Scholars with strong backgrounds in theological interactions with the sciences have waded into this discussion (e.g., Ronald Cole-Turner, Ted Peters, and Ilia Delio, and others), and several conversations can be found in related journals such as *Zygon* and *Theology & Science.*[7] The most impactful theologian to date in this area may be Noreen Herzfeld, who, from her first book-length reflection in 2002 to a special issue of the journal *Religions* in 2017, has continued to prod the depths of technology's impact on theological anthropology and expand the discussion to an increasing number of theologians.[8]

The second major area of theological reflection is within theological ethics. On this front there has been even less work, although an uptick can be seen in the past few years, particularly in survey essays that seek to bring more theologians to the field, as I am doing partially here. I am grateful for the work of Derek Schuurman, Brian Patrick Green, and Beth Singler in the past five years, whose surveys of the field prove useful in mapping the possibilities of theological engagement.[9] Green's introduction of *Laudato Sí'* as a framework for Catholic ethical reflection on technology is particularly helpful in laying foundations for a comprehensive Catholic theological ethic for technology, especially when combined with powerful applications of *Laudato Sí'* onto the digital age, such as Brianne Jacobs's recent work on Google and the technocratic paradigm.[10]

---

[7] Ronald Cole-Turner, "The Singularity and the Rapture: Transhumanist and Popular Christian Views of the Future," *Zygon* 47, no. 4 (2012): 777–96; Ronald Cole-Turner, *Transhumanism and Transcendence: Christian Hope in an Age of Technological Enhancement* (Washington: Georgetown University Press, 2011); Ilia Delio, OSF, "Artificial Intelligence and Christian Salvation: Compatibility or Competition?," *New Theology Review* 16 (2013): 39–51; and Brent Waters, *From Human to Posthuman: Christian Theology and Technology in a Postmodern World* (London: Routledge, 2016).

[8] Noreen Herzfeld, *In Our Image: Artificial Intelligence and the Human Spirit* (Minneapolis: Fortress, 2002); and Noreen Herzfeld, "Introduction: Religion and the New Technologies," *Religions* 8, no. 7 (2017): 129, doi.org/10.3390/rel8070129.

[9] Brian Patrick Green, "The Catholic Church and Technological Progress: Past, Present, and Future," *Religions* 8, no. 6 (June 2017): 106, doi.org/10.3390/rel8060106; Derek C. Schuurman, "Artificial Intelligence: Discerning a Christian Response," *Perspectives on Science and Christian Faith* 71, no. 2 (2019): 75–82; and Beth Singler, "An Introduction to Artificial Intelligence and Religion for the Religious Studies Scholar," *Implicit Religion* 20, no. 3 (2017): 215–31.

[10] See Brian Patrick Green, "Ethical Reflections on Artificial Intelligence," *Scientia et Fides* 6, no. 2 (October 9, 2018): 9–31, doi.org/10.12775/SetF.2018.015; and Brianne Jacobs, "Personhood, Bodies, and History in Google's Manifestation of the Technocratic Paradigm," in *Integral Ecology for a More Sustainable World: Dialogues with Laudato Sí'*, ed. Dennis O'Hara, Matthew Eaton, and Michael Ross (New York: Lexington, 2019), 221–34.

Given the limited number of theological reflections on technology ethics in general and the recent emergence of what I call ethics of technology and society, it should not be surprising that there is little theological reflection directly on the subfield, or hardly any surveys of this subfield beyond minor inclusions in the works above.[11] Before offering my own contribution to this discussion, I would like to expand briefly upon some new studies within ETS, focusing on the previously mentioned concentrations of access, representation, and structural biases.

ETHICS OF TECHNOLOGY AND SOCIETY: ADDRESSING BIAS, INEQUITY, AND RACISM

I identify three main concentrations within the last two decades of work in the subfield of ETS: access, representation, and structural biases. These questions and problems each have their own bibliographies, key players, and policy initiatives, but are tied together in their shared reflection on technology's relationship to ethical issues of bias, inequity, and racism.[12] The following discussion will draw attention to key figures and issues in each concentration, focusing more on representation and structural biases than on questions of access, for reasons that will become clear.

In the early development of the internet, sociologist Ruha Benjamin writes, "Much of the early research and commentary on race and information technologies coalesced around the idea of the 'digital divide,' with a focus on unequal access to computers and the [i]nternet that falls along predictable racial, class, and gender lines."[13] As such, many proposed solutions to the problem of unequal access are themselves riddled with biases, assumptions, and savior complexes of a techno-utopia. In nearly every instance of techno-utopia, writes Alondra Nelson, "racial identity, and blackness in particular" becomes "the anti-avatar of digital life. Blackness gets constructed as always oppositional to technologically driven chronicles of progress. That race (and gender) distinctions would be eliminated with technology

---

[11] See especially, Green, "Catholic Church and Technological Progress," 1–6.

[12] In this essay, I differentiate these terms as follows: bias defines intentional or unintentional preferential treatment of one group of people over another for any reason whatsoever (e.g., women, Jews, immigrants). Inequity defines any state of inequality which can be defined as unfair or unjust. For example, economic inequality could define the macroeconomic imbalance of wealth towards some countries or some individuals, whereas economic inequity would name this inequality as a systemic injustice. Both terms are differentiated from racism, which, following Matthew Clair, Jeffrey Denis, and W. J. Wilson, I define as "an ideology of racial domination," different from both racial discrimination (a bias) and racial inequality (an inequity). See Matthew Clair and Jeffrey S. Denis, "Sociology of Racism," in *The International Encyclopedia of the Social and Behavioral Sciences,* ed. James D. Wright (Amsterdam: Elsevier, 2015), 19:857–63.

[13] Benjamin, *Race After Technology,* 41–42.

was perhaps the founding fiction of the digital age."[14] Access discussions too often eliminate, rather than celebrate, difference, minimizing the actual, vital conversations needed around providing safe, affordable, and fast internet access to many people, a problem accentuated during the Covid-19 pandemic. As the pandemic worsened nearly every measurable aspect of inequality, policy changes concerning technology access are interwoven with policy discussions around childcare, taxation, wages, health care, and education.

Gender and racial representation among hiring practices at tech companies has been an issue in the tech sector since its origin. Discussions of representation, write scholars from the *AI Now* nonprofit, are "about gender, race, and most fundamentally, about power." Diversity affects "how AI companies work, what products get built, who they are designed to serve, and who benefits from their development."[15] In the past ten years, even as companies have become more aware of the problematic nature of discriminatory hiring processes, progress has been slow. According to a 2019 report from *Wired,* four of the major tech companies (Apple, Facebook, Google, and Microsoft) each reported less than 6 percent of Black Americans in their workforce, less than half the representative 13 percent population of Black Americans.[16] These numbers seem consistent throughout the technology sector, with Pew Research reporting 7 percent Black workers in the general computer industry.[17] The same study shows that women "remain underrepresented" in computer/tech occupations, with the percentage of women actually decreasing from 30 percent to 25 percent from 2000 to 2019.[18]

Discussions of diversity within the field of technology reveal many underlying cultural biases, but the least obvious and most insidious in its subtlety may be the primacy of meritocracy. "Studies have shown that a belief in your own personal objectivity, or a belief that you are not sexist, makes you less objective and more likely to behave in a sexist way," writes Caroline Criado Perez in *Invisible Women: Data*

---

[14] Alondra Nelson, "Introduction: Future Texts," *Social Text* 20, no. 2 (2002): 1.

[15] S. M. West, M. Whittaker, and K. Crawford, "Discriminating Systems: Gender, Race, and Power in AI," *AI Now Institute*, April 2019, ainowinstitute.org/discriminatingsystems.html.

[16] Sara Harrison, "Five Years of Tech Diversity Reports—and Little Progress," *Wired*, October 1, 2019, www.wired.com/story/five-years-tech-diversity-reports-little-progress/.

[17] Brian Kennedy, Richard Fry, and Cary Funk, "6 Facts about America's STEM Workforce and Those Training for It," *Pew Research Center*, April 14, 2021, www.pewresearch.org/fact-tank/2021/04/14/6-facts-about-americas-stem-workforce-and-those-training-for-it/.

[18] Kennedy, Fry, and Funk, "6 Facts about America's STEM Workforce and Those Training for It."

Bias in a World Designed for Men.[19] "Men who believe that they are objective in hiring decisions are more likely to hire a male applicant than an identically described female applicant. And in organizations which are explicitly presented as meritocratic, managers favor male employees over equally qualified female employees."[20] The belief in pure meritocracy decreases the propensity for objectivity in qualified hiring practices, revealing the idea of meritocracy as it is: a philosophy heavily influenced by misogynistic, ableist, white supremacist notions of intelligence, ambition, and social norms, largely utilized by people thinking themselves objective to perpetuate systems of inequity.[21]

The intersecting issues of access to technology and representation within the tech industry significantly contribute to and are affected by the third major concentration of technology and society studies: structural bias in technology itself. The very idea of structural bias in technology is deeply related to assumptions of objectivity in math and sciences, including the notion of meritocracy. The question of structural bias in technological development rejects deterministic theories of technological progress that allow discussions of access and diversity, as well as all ethical discussions of applied technology, but insist that the technology itself is objective. In this view, argues Benjamin, "Technology is often depicted as neutral, or as a blank state developed outside political and social contexts, with the potential to be shaped and governed" like any tool "through human action."[22] These deterministic and progressive technological philosophies reject any notion of social influence on the development of technology itself. Such philosophies have been widely rejected both throughout the wider field of science and technology studies since its origin (e.g., Jacques Ellul, *The Technological Society*), as well as throughout its intellectual offspring, the subfield of ethics discussed here. Both fields describe and analyze the deeply entangled ways in which culture, humanity, and identity have forever been transformed, and will continue to be transformed, by modern technology. "Technology is society," writes Manuel Castells in *The Rise of the Network Society* in 2009, "and society

---

[19] Caroline Criado Perez, *Invisible Women: Data Bias in a World Designed for Men* (New York: Abrams, 2019), "The Myth of Meritocracy"; Eric Luis Uhlmann and Geoffrey Cohen, "'I Think It, Therefore It's True": Effects of Self-Perceived Objectivity on Hiring Discrimination," *Organizational Behavior and Human Decision Processes* 104, no. 2 (2007): 207–23.

[20] Perez, *Invisible Women*.

[21] See Mar Hicks, "Meritocracy and Feminization in Conflict: Computerization in the British Government," in *Gender Codes: Why Women Are Leaving Computing,* ed. Thomas Misa (Hoboken, NJ: IEEE-CS/Wiley, 2010); Daniel Markovits, *The Meritocracy Trap: How America's Foundational Myth Feeds Inequality, Dismantles the Middle Class, and Devours the Elite* (New York: Penguin, 2019); and Michael J. Sandel, *The Tyranny of Merit: What's Become of the Common Good?* (New York: Farrar, Straus, and Giroux, 2020).

[22] Benjamin, *Race After Technology,* 41.

cannot be understood or represented without its technological tools."[23] Society's inherent biases, affectations, and desires are forever intertwined with the development of technology and science.

This argument has opened the door to several areas of new research within ETS, including algorithmic development biases, dataset biases, production biases, default statuses of (almost always) white men, objectivity biases, and confirmation biases. Several excellent examples of this research include works by Ruha Benjamin, Cristina Perez, Cathy O'Neill, and Kate Crawford. Benjamin's work brings together critical race studies with science and technology studies to examine the role of bias and racism throughout technological systems, from biometric technology to DNA tracking to policing to search engines. In the aforementioned book *Invisible Women,* Perez reflects upon the myriad ways that datasets and algorithms allocate resources inequitably by treating men as "standard" and women as "atypical."[24] In Cathy O'Neil's *Weapons of Math Destruction*, she explains that Big Data solutions are almost always flawed, perpetuating and often exacerbating inequality, by examining systems like insurance, policing, college admissions, and job applications.[25] Finally, Kate Crawford's 2021 *Atlas of AI* analyzes technological systems as the physical products they are, utilizing vast amounts of minerals, energy, labor, space, political power, and secrecy, and requiring ethical investigations into every aspect:

> Artificial intelligence is both embodied and material, made from natural resources, fuel, human labor, infrastructures, logistics, histories, and classifications. AI systems are not autonomous, rational, or able to discern anything without extensive, computationally intensive training with large datasets or predefined rules and rewards. In fact, artificial intelligence as we know it depends entirely on a much wider set of political and social structures. And due to the capital required to build AI at scale and the ways of seeing that it optimizes AI systems are ultimately designed to serve existing dominant interests. In this sense, artificial intelligence is a registry of power.[26]

Crawford's descriptions of AI can serve a metonymic function in relationship to technological development as a whole, establishing parameters and hermeneutics through which the entire field of tech is wrenched from its false objectivity, its male-dominated systems, and its belief that technology will solve all the problems that humanity

---

[23] Benjamin, *Race After Technology,* 41; Manuel Castells, *The Rise of the Network Society,* 2nd ed. (Oxford: Wiley Blackwell, 2009), 5.
[24] Perez, *Invisible Women,* "Introduction."
[25] Cathy O'Neil, *Weapons of Math Destruction* (New York: Crown, 2016).
[26] Katie Crawford, *Atlas of AI* (New Haven: Yale University Press, 2021), "Introduction," 8.

cannot. For example, in the process of examining labor practices, Crawford explicitly does not engage with robotic replacement debates, but focuses instead on "how humans are increasingly treated like robots and what this means for the role of labor."[27]

While many more examples of this subfield exist,[28] the above texts exemplify an approach to technology and society that should resonate with theological scholars. I found particular resonance with the work of M. Shawn Copeland as I came to know this subfield. In the following sections of this essay, I will attempt to establish a novel theological dialectic using Copeland's employment of Lonergan and Metz in her mystical-political theology of the common, human good.

A COMMON, HUMAN GOOD

In 2019, at a conference called "The Common Good in the Digital Age," Pope Francis commended the participants for working to bridge the gap between technological development and the common good:

> If technological advancement became the cause of increasingly evident inequalities, it would not be true and real progress. If humanity's so-called technological progress were to become an enemy of the common good, this would lead to an unfortunate regression to a form of barbarism dictated by the law of the strongest….A better world is possible thanks to technological progress, if this is accompanied by an ethic inspired by a vision of the common good, an ethic of freedom, responsibility and fraternity, capable of fostering the full development of people in relation to others and to the whole of creation.[29]

---

[27] Crawford, *Atlas of AI*, "Two: Labor."

[28] E.g., Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the World* (Cambridge: MIT Press, 2018); Simone Browne, *Dark Matters: On the Surveillance of Blackness* (Durham, NC: Duke University Press, 2015); Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Conference on Fairness, Accountability, and Transparency* (2018), 77–91; Wendy Hui Kyong Chun, "Introduction: Race and/as Technology; or How to Do Things to Race," *Camera Obscura: Feminism, Culture, and Media Studies* 24, no. 1 (70) (2009): 7–35; Jessie Daniels, *Cyber Racism: White Supremacy Online and the New Attack on Civil Rights* (Lanham, MD: Rowman & Littlefield, 2009); Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York: St. Martin's, 2018); Mar Hicks, *Programmed Inequality: How Britain Discarded Women Technologists and Lost its Edge in Computing* (Cambridge: MIT Press, 2017); and Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru: "Detecting Bias with Generative Counterfactual Face Attribute Augmentation," *arXiv preprint arXiv:1906.06439* (2019).

[29] Pope Francis, "To the Participants in the Seminar 'The Common Good in the Digital Age,' Organized by the Dicastery for Promoting Integral Human Development (DPIHD) and the Pontifical Council for Culture (PCC)," September 27, 2019, www.vatican.va/content/francesco/en/speeches/2019/september/documents/papa-francesco_20190927_eradigitale.html.

The cautious technological optimism that Pope Francis displays here is founded upon his vision of the *common good*, a phrase with centuries of philosophical and ethical tradition and theological interpretation.[30] As an ethic of solidarity, community, and accountability, it can be a powerful rhetorical device upon which to build consensus, but given its wide usage, the phrase alone requires further explanation in order to carry any demonstrable weight. The *Catechism of the Catholic Church* casts the idea of the human good as one rooted in individual human dignity, community development, peace, justice, stability, and progress, and one that has become a treasured inheritance from ancient Christian traditions (nos. 1905–12). Despite its use in discussions of tech ethics, including in a wide range of secular and ecclesial ethical guidelines,[31] the common good is not a phrase that seems to add new insight into contemporary discussions of technological ethics, especially for scholars of technology and society.

In order to understand and address this situation, in which a powerful ethical idea has lost its value, I will now examine M. Shawn Copeland's utilization of Bernard Lonergan's idea of the human (common) good. Lonergan worked throughout his life to develop a more concrete vision of the common good, which he termed the "human good." Lonergan's human good was a central part of his overall systematic theology, representing not an eschatological unobtainable goal but a "comprehensive, and hence not abstract" goal for human society.[32] The human good "is not a system, a legal system or a moral system. It is a history, a concrete, cumulative process resulting from developing human apprehension and human choices that may be good or evil."[33] The proceeding analysis will examine a transformation of Copeland's understanding of Lonergan's human good in order to develop a framework of mystical-political theology. This framework will then connect with the works of Benjamin, O'Neil, Perez, and Crawford, who serve as exemplars through which an application of

---

[30] Green articulates the nuanced path that Francis tries to walk both here and in *Laudato Si':* "For nearly its entire history the Church has stood for the preservation and advancement of knowledge and technology, with exceptions only for a few of those technologies [e.g., weapons of mass destruction, embryonic stem cell research, environmentally unsustainable technologies] which it evaluates as preventing or harming human life. *Laudato Si'* is best interpreted in light of this tradition" (Green, "The Catholic Church and Technological Progress," 9).

[31] Pontifical Academy for Life, "The Call—Rome Call For AI Ethics," www.romecall.org/the-call/. For a good summary of commonalities and differences between various ethical guidelines, see Anna Jobin, Marcello Ienca, and Effy Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence* 1, no. 9 (2019): 389–99; Berkman Klein Center, "Principled Artificial Intelligence," February 5, 2020, cyber.harvard.edu/publication/2020/principled-ai.

[32] Bernard Lonergan, *Topics in Education: The Cincinnati Lectures of 1959 on the Philosophy of Education* (Toronto: University of Toronto Press, 1993), 28.

[33] Lonergan, *Topics in Education*, 33.

Copeland's mystical-political theological dialectic can be discerned and developed.

THE PRACTICAL HUMAN GOOD

M. Shawn Copeland has been a student of Bernard Lonergan's work since her doctoral studies in the 1980s. She is one of the leading interlocutors of Lonergan's theories in the 21st century, and one of the foremost Catholic theologians in the world, best known for her work within womanist Catholic theology. To understand how Copeland interacts with Lonergan's concept of the human good, one must begin with Copeland's understanding of four concepts central to Lonergan's work: horizon, bias, progress, and decline. "Horizon," writes Copeland, "connotes a worldview" and "bias may participate in the construction and control of it, but both govern meaning-making."[34] Horizon is what we see; bias distorts that vision by disrupting our judgment, intelligence, common sense, and sense of what makes community.[35] Our present existence, then, consists of a personal and social horizon constantly inhibited by various forms of bias. How we move beyond this present condition is described by progress and decline within an overall matrix termed *human good.*

At first glance, the concepts are self-explanatory. We aim and search for the overall human (common) good. We experience progress and decline in this search. Progress is the lessening of biases; decline is the increase of biases. For Lonergan and Copeland, though, the human good is not only the endpoint but the structure within which everything occurs. Progress is not merely positive movement but the example of "persons struggling to live attentively, intelligently, rationally, and responsibly." Decline exists as a result of "oversight, inattention, unreasonableness, and irresponsibility."[36]

For Lonergan, bias, horizon, progress, and decline are necessary to form the methodological framework of the human good. This brings us to the first definition of human good for Copeland: the human good is a *structure* which allows for interaction between religion and the "cultural matrix" of society.[37] To be effective, this structure of the

---

[34] M. Shawn Copeland, *Enfleshing Freedom: Body, Race, and Being* (Fortress, Minneapolis: 2010), 9.

[35] These four facets of bias are properly considered dramatic (judgment), individual (intelligence), general (common sense), and group (community). Copeland describes them numerous times throughout her corpus, including in *Enfleshing Freedom*, 12–15.

[36] M. Shawn Copeland, "The Interaction of Racism, Sexism, and Classism in Women's Exploitation," in *Women, Work, and Poverty*, ed. Elisabeth Schüssler Fiorenza, Anne E. Carr, and Marcus Lefébure (Edinburgh: T. & T. Clark, 1987), 20.

[37] "If theology mediates between a cultural matrix and the significance and role of religion in that matrix, then the theologian needs some framework by which to attend concretely to the cultural matrix as it is in process" (Copeland, "The Interaction of Racism," 19).

human good must anticipate complexities as well as offer explanatory accounts of meaning in the relationship between religion and the social matrix. As such, the human good "charts progress and change as well as decline and breakdown" by focusing on three areas of interaction with society: "(a) individuals in their potentialities and actuations, (b) cooperating groups, and (c) the ends, the values by and for which individuals and groups act."[38] In each of these groups, the human good is "a field theory" and "a set of fixed terms" which allows us to understand how bias influences our horizons, which in turn allows us to see whether we are achieving progress or decline.[39]

This concrete structural matrix of cultural interaction retains significant power for Copeland's theological approaches to real instances of bias and oppression. Her early reflections on Black theology, for example, indicate the importance of this structure: "As a politically responsible methodical theology, black theology mediates the significance of religion within a cultural matrix. Black theology as politically responsible methodical theology must apprehend and understand the social and cultural matrix in which it seeks to mediate Christian religion. The structure of the human good provides a way for the theologian to think concretely about that matrix, since the structure is the form of society."[40] In this definition of the human good, Copeland finds an effective "heuristic structure or implicit definition or field theory for apprehending, criticizing, and evaluating the objective components of the social order and for inquiring into the condition of the human good which is interchangeable with human history. Neither an abstraction nor a utopian ideal, the human good is the concrete, cumulative process resulting from the development of human apprehension and choice, from the integrated completion of various moments and stages of human potentiality."[41]

In this structure, rooted in practicality, Copeland sees "the basic terms for a political theology" since "Lonergan's theology offers the appropriate locus for the integration of the empirical human sciences in their approach to the problems that pervade the social order, because sin is manifest in the concrete human situation, with concrete results that can be disclosed as crime, as aberration, as an evil component in the social progress."[42] Crawford's descriptions of human rights abuses in the development of AI, for example, would constitute an ideal constructive use of human sciences, and this essay could be an example of theological integration therein.

---

[38] Copeland, "The Interaction of Racism," 20.
[39] Copeland, "The Interaction of Racism," 20.
[40] M. Shawn Copeland, "A Genetic Study of the Idea of the Human Good in the Thought of Bernard Lonergan," (Doctor of Philosophy Thesis, Boston College, 1991), 290.
[41] Copeland, "A Genetic Study," 197.
[42] Copeland, "A Genetic Study," 198.

This application of the human good has immediate import for Copeland, who employs Lonergan in her understanding of womanist theology. It is the "cognitive praxis" of enslaved Black women, Copeland writes in 1993, that formed their narratives and serves as the basis for a theology of suffering.[43] "Womanist theology claims the experiences of Black women as *proper and serious data* for theological reflection. Its aim is to elucidate the differentiated range and interconnections of Black women's gender, racial-ethnic, cultural, religious, and social (i.e., political, economic, and technological) oppression."[44] Since illumination of biases brings progress and self-transcendence on our path within and towards the human good, womanist theology offers "proper and serious data" in order to achieve this progress. "Only by attending to Black women's feelings and experiences, understanding and reflection, judgment and evaluation about their situation, can we adequately challenge the stereotypes about Black women— especially those stereotypes that coalesce around that most popular social convention of female sexuality, the 'cult of true womanhood.'"[45]

Several years later, in laying out a feminist theological solidarity as praxis, Copeland argues that the possibility for individual and social progress lay in our ability to "be attentive, to be intelligent, to be rational, and to be responsible."[46] Furthermore, Copeland's balance between the eschatological and the possible plays a direct role in her solidaristic conclusions. "By focusing on solidarity as a theological category," she writes, "I have hoped to call attention to the gap between rhetoric and Christian social praxis in expressions of feminist theology. Moreover, I have hoped to encourage diffuse, halting, yet, unfulfilled efforts toward a critical Christian feminist theology that aims for 'the basic transformation of [the whole of] society: a new order, not a new deal … [but] …a *new humanity.*'"[47]

THE MYSTICAL-POLITICAL HUMAN GOOD

In Copeland's dissertation, written in 1987 and focused on Lonergan's ideas of the human good, she notes that the practical, structural definition of human good fails to consider Lonergan's own description of the self-transcendent nature of the human subject. By tracing the

---

[43] M. Shawn Copeland, "Wading Through Many Sorrows: Toward a Theology of Suffering in Womanist Perspective," in *A Troubling in My Soul: Womanist Perspectives on Evil and Suffering*, ed. Emilie M. Townes (Maryknoll, NY: Orbis, 1993): 123–24.
[44] Copeland, "Wading Through Many Sorrows," 111. Emphasis added.
[45] Copeland, "Wading Through Many Sorrows," 111.
[46] M. Shawn Copeland, "Toward a Critical Christian Feminist Theology of Solidarity," in *Women and Theology*, ed. Mary Ann Hinsdale and Phyllis H. Kaminsky (Maryknoll, NY: Orbis, 1995), 23.
[47] Copeland, "Toward a Critical Christian Feminist Theology of Solidarity," 32–33. See Beatriz Melano Couch, "Statement," in *Theology in the Americas,* ed. Sergio Torres and John Eagelson (Maryknoll, NY: Orbis, 1976), 374.

account of the concept in Lonergan's *Method in Theology*, Copeland begins to develop another, overlapping definition for the human good as a "transcultural and transhistorical structure within which solutions to the problems of human living are worked out…. The standard of the human good is a complete life of authentic self-transcendence—the real life of good women and good men, authentic self-transcending subjects."[48]

This "transhistorical" vision of the human good played only a small role in Copeland's practical structure until, arguably, her address to the Catholic Theological Society of America in 1998.[49] There, Copeland begins to explore Lonergan's "Mystical Body of Christ" as a way in which to locate individual transcendence while still working towards the overall human good. "The Mystical Body of Christ is … not a theology; it is a 'divine solidarity in grace.' That solidarity makes a claim on each of us and a claim on theology: It obliges each of us to a social praxis in the here and now that resists the destructive deformation of sin in ourselves and in our society."[50] For Christopher Pramuk, this 1998 address marked a turning point in Copeland's theology and the beginning of Copeland's integration of Johann Baptist Metz's categories of mystical-political praxis and solidarity. Copeland's reflections on the mystical, he writes, were "never just mystical but always mystical-political, never triumphal but always rooted in 'the anguish of the victims.'"[51]

Copeland constructs this mystical vision alongside discussions of the practical, as can be seen in her essay on racism and Christian vocation from 2002: "Inasmuch as that determination is to be made before the cross of Christ, our theology must stand with society's most abject, despised, and oppressed. In this posture, our theology must repudiate the principalities and powers of society and resist their efforts to seduce its spirit-filled, prophetic, critical, and creative impulse."[52] Emphasizing the mystical, she adds that "only from rootedness in prayer and a desire for God and life in God can our theology elucidate

---

[48] Copeland, "A Genetic Study," 261.

[49] M. Shawn Copeland, "The New Anthropological Subject at the Heart of the Mystical Body of Christ," *Proceedings of the Catholic Theological Society of America* 53 (1998): 25–47.

[50] Copeland, "The New Anthropological Subject," 47. Bernard Lonergan, "Finality, Love, Marriage," in *Collection: Papers by Bernard Lonergan, S.J.*, ed. Frederick E. Crowe (Montreal: Palm, 1967): 26.

[51] See Christopher Pramuk, "'Living in the Master's House': Race and Rhetoric in the Theology of M. Shawn Copeland," *Horizons* 32, no. 2 (2005): 315. See Johannes Baptist Metz, *A Passion for God: The Mystical-Political Dimension of Christianity*, trans. J. Matthew Ashley (New York: Paulist, 1998).

[52] M. Shawn Copeland, "Racism and the Vocation of the Christian Theologian," *Spiritus* 2 (2002): 22.

a new and redemptive solidarity in the transforming reality that is Christ."[53]

From the late 1990s onward, the mystical-political framework grounds Copeland's descriptions of the human good in a political praxis of solidarity situated in "self-transcendence or being-in-love-with-God," understood through Jesus's example and his call to carry our cross and follow.[54] "Christian discipleship," she argues in 2003, "as a lived mystical-political way forms the locus for the fundamental grasp of who Jesus of Nazareth is and what following and believing in him means."[55] This framework leads her to espouse a new vision for political theology more generally, presented as her presidential address to the CTSA in 2004:

> Our political theology recognizes that life is vested with an "apocalyptic goal," which orients the horizon of our expectation toward the coming of the Lord; yet that orientation never surrenders its cultural and social responsibilities. Hence, political theology will scrutinize from the perspective of the excluded, despised, and poor, the development, promotion, and advance of programs and schemes that propose to resolve violence, injustice, and oppression. Further, political theology will provide a critique of the Church whenever it attempts to evade the dangerous memory of the crucified Jesus by slipping into what Metz names a "fatal banality" or an irenic conformity so passive that it glides over the resolute work of authentic peace, thereby betraying its mystery.[56]

As Copeland turns toward Christological embodiment and begins to engage with Metz's notion of mystical-political discipleship, Lonergan's framework of the human good subtly shifts. Note the difference between her piece on the human good in 1987 and her presidential address in 2004:

> If theology mediates between a cultural matrix and the significance and role of religion in that matrix, then the theologian needs some framework by which to attend concretely to the cultural matrix as it is in process….

---

[53] Copeland, "Racism and the Vocation," 27.

[54] M. Shawn Copeland, "Knowing Christ Crucified: Dark Wisdom from the Slaves," in *Missing God?: Cultural Amnesia and Political Theology (Festschrift for Johann Baptist Metz)*, ed. John K. Downey, Steven T. Ostovich, and Jürgen Manemann (Berlin: LIT, 2006), 60.

[55] M. Shawn Copeland, "The Cross of Christ and Discipleship," in *Thinking of Christ: Proclamation, Explanation, Meaning*, ed. Tatha Wiley (New York: Continuum, 2003), 179.

[56] M. Shawn Copeland, "Political Theology as Interruptive," *Proceedings of the Catholic Theological Society of America* 59 (2004): 79; Johann Baptist Metz, *Love's Strategy: The Political Theology of Johann Baptist Metz,* ed. John K. Downey (Harrisburg PA: Trinity International, 1999), 150.

> Such an instrument is provided by Bernard Lonergan's concept of the human good. (1987)[57]

> If a function of theology is "to mediate between a cultural matrix and the significance and role of religion in that matrix," then political theology constitutes a crucial, even necessary, framework for doing theology in our time, in the United States. (2004)[58]

"Political theology" replaces Lonergan's "human good" in describing the overall framework that Copeland desires for the function of theology. Of course, Copeland is not replacing Lonergan: the 2004 address is actually dedicated to Bernard Lonergan, "my teacher and yours."[59] Copeland continues to work constructively within the practical structure of Lonergan's human good, including an essay also published in 2004 that speaks powerfully of the need for the transformation of industry in Detroit via the framework of the human good.[60]

Nevertheless, the transformation from a concept of human good which *subsumes political aspects of theology* to a concept of human good which *is subsumed within a mystical-political theology* speaks to a significant maturation of Copeland's thought. To be faithful to Copeland's loyalty to Lonergan, it perhaps signals Copeland's ability to be more creative with Lonergan's thought in order to make room for a theological hermeneutic that might better address the needs of contemporary society and contemporary theology.

FORGIVENESS, WITNESS, MEMORY, AND LAMENT

In the search for a unique theological voice in the ethics of technology and society, one must look past the usual places in this diverse and highly charged ethical discussion. Condemnations of bias in machine learning, technological development, hiring practices, and digital access are necessary and ethical, but not uniquely theological. For example, the works by Benjamin, Crawford, Perez, and O'Neil argue strongly for ethical principles such as those affirmed by the Vatican's "Rome Call for AI Ethics": transparency, inclusion, responsibility, impartiality, reliability, security, and privacy.[61]

---

[57] Copeland, "The Interaction of Racism," 19.
[58] Copeland, "Political Theology as Interruptive," 72.
[59] Copeland, "Political Theology as Interruptive," 71.
[60] M. Shawn Copeland, "A Theologian in the Factory: Toward a Theology of Social Transformation in the United States," in *Spirit in the Cities: Searching for Soul in the Urban Landscape*, ed. Kathryn Tanner (Minneapolis: Fortress, 2004), 20–46.
[61] "The Rome Call for AI Ethics," Pontifical Academy for Life, www.romecall.org/the-call/. As I will discuss later, these principles are themselves vague and heavily corporatized, relying on previously established ethical frameworks from tech companies like Microsoft rather than originating values from the *Catechism*. The Rome Call, for example, is nearly identical to the previously published "Responsible

Furthermore, calls for employment of the common good, while enjoying a rich heritage within the Christian tradition, do not hold a unique space for theological voices in the technological ethics community, as such ethical guidelines and principles have been wholly incorporated into secular spaces.[62] Similar to calls against bias, this does not mean that arguments for the common good from Christian ethics are pointless—on the contrary! Following the lead of the Vatican, other Christian leaders must speak out strongly in favor of ethical principles throughout the tech industry. Such calls will likely bolster better conversations and actions among faithful Christians, but I fear they may have a small impact in the community of technological ethics.

In searching for a unique voice for theology within this growing field of technological ethics, I found Copeland's transformation of the practical, structural human good revelatory. Her mystical-political framework holds together related but easily disjointed strains within ethical teaching: the difficult, rational, practical development of the human good and the individual self-transcendence of being known and being in love with God. This duality—this both/and—was clearly a desired articulation for Copeland for a while, as it is presented in a nascent form in one of her earliest discussions of the power and promise of Black theology. "As politically responsible and methodical, black theology stands as a higher viewpoint which can reinforce the social scientist's detached, disinterested, unrestricted desire to know; it urges the social scientist to seek concrete practically intelligent and reasonable solutions to human problems. It calls the social scientist to put his or her intellectual efforts to the service of the progress of the common human good, to assume responsibility for creative and healing solutions to those problems even when the situation seems most opaque."[63] She continues, asking the social scientist—and by extension all who work on the critical discovery of social inequities—to examine their own self-transcendence:

> Moreover, a politically responsible methodical black theology proposes that the social scientist advert to his or her own interiority. The theologian poses to the social scientist sustained engagement with the very same questions with which he or she is committed to wrestle: What does it mean to know? … Do I know what it means to respect others, to be in love with them? Do I know what it means to be a human person? ... Do I know concretely what self-transcending love means?... Do I know what it means to suffer? Do I know what it means

AI Principles" from Microsoft. Microsoft, "Responsible AI," www.microsoft.com/en-us/ai/responsible-ai.

[62] Jobin, Ienca, and Vayena, "The Global Landscape of AI Ethics Guidelines"; Berkman Klein Center, "Principled Artificial Intelligence."

[63] Copeland, "A Genetic Study," 290.

> to be vanquished, to be colonized, to be a victim? Do I know what it means to be privileged, to be a colonizer, to be a victor?[64]

Two decades later, as she presented her call for an interruptive mystical-political theology to the members of the CTSA, she called all theology—not just Black theology—to be political, interruptive, and anti-oppressive. In doing so, she expounded a vision of mystical self-transcendence rooted in political theology, employing categories of forgiveness, reconciliation, memory, lament, and witness.[65]

Each category offers a modern approach from an ancient faith to technocratic systems that wield increasingly alarming levels of power and contribute unhelpfully to the state of violence, bias, and inequity. Forgiveness and reconciliation ground the initial approach to the modern world: they are neither "abstract concepts, nor mere emotion or feeling." For the followers of Jesus, Copeland writes, "The only appropriate responses to violence and malevolence are forgiveness and reconciliation."[66] Following this grounding, Copeland challenges each of us to confront inequities through witness, memory, and lament.

To witness is to tell the truth in a world living in falsehood. "The martyr witnessed for her or his faith even if that witness involved self-sacrifice or death….The witness is never a spectator, never a dilettante. In order to interrupt the violence that tears at the fabric of our society, in order to do political theology, we theologians must be willing to sacrifice—our comforts, our security, our joys, perhaps, our lives."[67] Witness leads to remembrance, to allowing ourselves dangerous memories. As a part of this witness, we must "recover and expose memories that we have been too fearful and too ashamed to admit and confront….We theologians must take seriously the 'negativity of history in its interruptive and catastrophic character,' for these histories of suffering form the theological locus of our truth-telling."[68]

Witness and memory, together, call us to lament. Part and parcel of this truth telling and recovery of memory, we must lead the community in a lament that "announces aloud and publicly what is unjust in the here-and-now." Lament, she continues, "protests, pushes against that calculus of power by which the weak and the vulnerable suffer oppression and abuse. Lament not only dialogues, but also boxes with God—questions, argues, and rebukes. In this way, lament takes seriously God's compassionate love and care in the midst of suffering and privation….Lament names and grieves injustice,…lament

---

[64] Copeland, "A Genetic Study," 291–92.
[65] Copeland, "Political Theology as Interruptive," 79–81.
[66] Copeland, "Political Theology as Interruptive," 79–80.
[67] Copeland, "Political Theology as Interruptive," 80.
[68] Copeland, "Political Theology as Interruptive," 81. See Metz, *Love's Strategy*, 150, 139.

names and grieves social pain…[and] lament makes 'spaces of recognition and catharsis' that prepare for justice."[69]

As forgiveness and reconciliation ground our initial focus of self-transcendence, lament grounds the space from which we must witness, speak the truth, and lift up the dangerous memories of the past, which both ground our present and determine our future. "Without pain brought to the open, seen, and heard, paid attention to and acknowledged," writes Kathleen O'Connor, genuine change of long-held biases, now appearing in technocratic systems of power and privilege, is impossible.[70]

Now, finally, we find a unique voice, a unique space, for theology. We begin with the common human good, with statements from the Vatican, bishops, theological ethicists, and governing bodies joining the chorus of secular ethicists and scholars of technology and society in arguing for things like autonomy, dignity, transparency, privacy, equity, diversity, and inclusion.[71] Following Lonergan's framework of praxis, we are attentive, intelligent, reasonable, and responsible, working with tech companies, developing policy, writing code, and informing the public. In the same breath, we break for the individual, for the community, for the world. We witness, we remember, we lament—actions which may come across as impractical, but which ground us as Church, as humans, as individuals before God. We give space in our churches and in our schools to witness, remember, and lament injustice, bias, and hate. We seek forgiveness for our complicities and seek reconciliation where it can be found. We find holiness in the suffering individual and the community. We name, remember, and lament the lives torn apart and lost:

1. The millions of people whose images are used without consent in facial recognition and biometric databases by governments and private industries;[72]
2. Countless people of color who have been unjustly arrested, harassed, incarcerated, and killed from the use of the policing algorithm PredPol;[73]
3. The millions of workers who labor in poor conditions without unions around the world, in order to generate massive quantities of wealth for the billionaire tech class;[74]

---

[69] Copeland, "Political Theology as Interruptive," 81. See Kathleen O'Connor, *Lamentations and the Tears of the World* (Maryknoll, NY: Orbis, 2002), 128.
[70] O'Connor, *Lamentations*, 132.
[71] A wonderful example of theological articulation of this is Brianne Jacobs's aforementioned arguments for labor transparency and human dignity, found in "Personhood, Bodies, and History," 230–32.
[72] Crawford, *Atlas of AI*, "Three: Data."
[73] Benjamin, *Race After Technology,* 80–87.
[74] Jay Greene, "Riots, Suicides, and Other Issues in Foxconn's iPhone Factories," *CNET*, September 25, 2012, www.cnet.com/news/riots-suicides-and-other-issues-in-

> 4. Countless women who met untimely deaths or suffered needlessly because most of the default medical data used in textbooks is, to this day, from male bodies;[75]
> 5. The children who labor in rare mineral mines around the world, for "to understand the business of AI, we must reckon with the war, famine, and death that mining brings with it";[76]
> 6. Black, Jewish, LGBTQ, Muslim, Asian, Asian-American, African, Latinx, and all individuals who have been targets of online extremism and hate, which has frequently turned into in-person violence, made so much easier through the non-regulation of social media;
> 7. The Earth itself, in a climate emergency, exacerbated by the relentless stripping of nonrenewable elements such as cobalt, lithium, nickel, and so many others at such a frenzied pace that tech tycoons are now spending billions attempting to escape the Earth;[77]

And then we break again for praxis, new policies, guidelines, apps, companies, and public awareness. The two movements sit apart and yet together, mystical and political, contemplation and action, prayer and work. It is the oldest both/and in the Christian tradition, and perhaps the most important.

The import of this dialectic is both individual and social, personal and corporate. Corporations must be held accountable for their historical and continued failings, including perpetuating inequities, insufficiently addressing bias, and ignoring racist practices and policies. We, the Church, must be active participants in framing the future of technology, and we must neither be compromised by the wealth nor lured into pronouncing toothless guidelines. For example, how does the Rome Call for AI Ethics ensure accountability to the practices so many corporations have now promised to uphold? Will there be ecclesial delegates that inspect algorithms, hiring practices, and use guidelines? If we allow tech leaders to take pictures with the Pope and sign a document without holding them accountable for inequitable practices, we fall victim to the same allure of power that tech giants wield effortlessly around the globe.

The Church has no responsibility to the modern giants of technology, but it holds a deep responsibility to those who suffer continued

---

foxconns-iphone-factories/; Michael Sainato, "'I'm Not a Robot': Amazon Workers Condemn Unsafe, Grueling Conditions at Warehouse," *The Guardian,* February 5, 2020, www.theguardian.com/technology/2020/feb/05/amazon-workers-protest-unsafe-grueling-conditions-warehouse.

[75] Perez, *Invisible Women,* "Part IV: Going to the Doctor."

[76] Crawford, *Atlas of AI,* "One: Earth"; Siddharth Kara, "Is Your Phone Tainted by the Misery of 35,000 Children in Congo's Mines?," *The Guardian*, October 12, 2018, www.theguardian.com/global-development/2018/oct/12/phone-misery-children-congo-cobalt-mines-drc.

[77] Crawford, *Atlas of AI*, "Coda: Space."

inequities through the modern technocratic paradigm, as Pope Francis himself has noted: "Science and technology are not neutral; from the beginning to the end of a process, various intentions and possibilities are in play and can take on distinct shapes. [We need] to appropriate the positive and sustainable progress which has been made, but also to recover the values and the great goals swept away by our unrestrained delusions of grandeur" (*Laudato Si'*, no. 114). We, as individuals, communities, scholars, and Church must bear witness to the loss, discover the bias, support the researchers who look for injustice, lift up hackers and coders whose hearts burn for equity and justice, and demand righteous policies and practices from corporations and governments that ensure the dignity, privacy, and security of all individuals. Only then, in the holy tradition of the mystical and political, can we find a unique, potent, and liberative theological voice. M

John P. Slattery, PhD, is a Senior Program Associate with the Dialogue on Science, Ethics, and Religion Program at the American Association for the Advancement of Science in Washington, DC, and a Fellow for the Grefenstette Center on Ethics in Science, Technology, and Law at Duquesne University. Slattery has published two recent volumes: *Faith and Science at Notre Dame: John Zahm, Evolution, and the Catholic Church* (Notre Dame Press, 2019), and *Christian Theology and the Modern Sciences* (edited, T. & T. Clark, 2020).

Journal of Moral Theology, Vol. 11, Special Issue 1 (2022): 132–156

Can a Robot Be a Person? De-*Facing* Personhood and Finding It Again with Levinas

Roberto Dell'Oro

De ce terrible paysage,
Tel que jamais mortel n'en vit,
Ce matin encore l'image,
Vague et lointaine, me ravit.

Le sommeil est plein de miracles!
Par un caprice singulier
J'avais banni de ces spectacles
Le végétal irrégulier,

Et, peintre fier de mon génie,
Je savourais dans mon tableau
L'enivrante monotonie
Du métal, du marbre et de l'eau.
(Baudelaire)[1]

THE QUESTION "CAN A ROBOT BE A PERSON?" has emerged of late in the field of bioethics. It is a fitting provocation, an instigation to think impelled by technological advances, as have been the many ethical issues posed by developments in medicine over the past century.[2] Robots can do almost everything:

[1] "This morning I am still entranced - By the image, distant and dim - Of that awe-inspiring landscape - Such as no mortal ever saw. Sleep is full of miracles! Obeying a curious whim, I had banned from that spectacle - Irregular vegetation - And, painter proud of his genius - I savored in my picture - The delightful monotony- Of water, marble, and metal" ("Parisian Dream," *The Flowers of Evil*, fleursdumal.org/poem/228).

[2] The emergence of robotics has triggered reflections at different levels, mostly with a concern for the epistemological and ethical dimensions of the impact of robots and artificial intelligence on human life. The anthropological reflection (in the sense of a *philosophical* anthropology) is less developed. A general introduction to the problem is found in Phil Husbands, "Robotics," in *Cambridge Handbook of Artificial Intelligence*, ed. K. Frankish and W. M. Ramsey (Cambridge: Cambridge University Press, 2014), 269–95; Patrick Lin, Keith Abney, and George A. Bekey, eds., *Robot Ethics: The Ethical and Social Implications of Robotics* (London: MIT Press, 2012); Margaret

discharge complex logical operations, resolve algorithmic puzzles impossible to the average human brain, carry out operations on command, and even act *of their own accord*, posing the issue of whether their endowments, either in the cognitive field or in the sphere of autonomous decision-making, might not make them closer to us than we think.[3] Thus the question: can a robot be a person?[4]

The question points to a doubt, a puzzlement about the nature of personhood with respect to its attributions. Our world has become the abode of *homo technologicus*, the impersonal space of the Neuter, as Levinas might put it: a world of objects exposed to the totalizing gaze of science, whose *outlook has become a doing*.[5]

The suggestions advanced by Japanese scientist Hirosi Ishiguro are telling. For years, Ishiguro has worked toward the creation of interactive robots, specifically, robots with a human appearance or *androids*, on the premise that "we empirically know the effect of appearance is

A. Bode, *The Philosophy of Artificial Intelligence* (Oxford: Oxford University Press, 1990). In a critical vein, Hubert L. Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (Cambridge, MA: MIT Press, 1994) and Guy Vallancien, *Homo Artificialis* (Paris: Michalon, 2017).

[3] Japanese scientist Hiroshi Ishiguro, possibly one of the most renowned in the field, speaks of "human-robot symbiotic society." See Hiroshi Ishiguro, "Studies on Interactive Humanoids," in *Robo-Ethics: Humans, Machines, and Health*, ed. Vincenzo Paglia and Renzo Pegoraro (Rome: Pontifical Academy for Life, 2020), 67–102.

[4] Although often robots, humanoids, and artificial intelligence are considered alike, this is not necessarily the case. To begin, one ought to distinguish between so called embodied and non-embodied machines. Furthermore, machines, whether embodied or non-embodied, may be provided with some form of artificial intelligence. They are either "stupid"—i.e., programmed to work automatically—or "intelligent"—that is, endowed with increasing cognitive and decisional abilities. Roberto Cingolani sums up the meaning of the distinctions in question thus: "The availability of increasingly powerful calculation machines is constantly extending the limitations of artificial intelligence. At the same time, it is allowing the development of increasingly performing embodied machines (provided with sight, touch, and biomechanical abilities), making realistic the assumption that robots are characterized by performances increasingly closer to those of the human" (Roberto Cingolani, "Robots and Intelligent/Autonomous Systems: Technology, Social Impact, and Open Issues," in Pegoraro and Paglia, *Roboethics*, 34).

[5] Though Levinas is critical of Heidegger, the latter has offered a powerful critique of technology (cybernetics) as the inevitable destiny of metaphysics. We dwell in the *Gestell*, in the "framework" provided by the totalizing outlook of technology and the forgetfulness of being. See Martin Heidegger, *The Question Concerning Technology and Other Essays*, trans. William Lovitt (New York: Harper, 1977), 14–17. Another interesting take on the relation of technology to modernity is Monette Vacquin, "The Monstrous as the Paradigm of Modernity? Or Frankenstein, Myth of the Birth of the Contemporary," *Diogenes* 49/3, no. 195 (2002): 27–33. For a general philosophical analysis of the phenomenon of technology, Don Ihde, *Technology and the Lifeworld* (Bloomington: Indiana University Press, 1990).

as significant as behaviors in communication."[6] To tackle the problem of appearance and behavior, two approaches are necessary: one from robotics, and the other from cognitive science. The cross-disciplinary framework emerging from such interaction is *android science*, whose goal is the creation of robots with humanlike appearance, movement, behavior, and perception. Although further study is needed to address meta-level cognitive functions that more intelligent human-friendly robots might perform (intelligence embodiment, multi-modal integration, intention/desire, consciousness, and social relationships), the goal of scientific research, according to Ishiguro, is to "develop companion robots that can pass the Total Turing Test as a scientific and engineering goal." The premise of the research in question is that the robot, having passed all the tests that evaluate its "total human-likeness," will be "accepted as a member of our society."[7]

The difficult question at stake here is "recognition." The robot is not the person, but its human-like appearance and behavior allows the *other* to the robot, potentially a conversation partner, to be fooled into thinking it is a person.[8] The relation between human and robot thus rests on a kind of game of pretense: though one knows the robot is not human, he/she can still deal with it as a social partner. This seems to be the hypothesis tested out experimentally by scientists like Ishiguro:

> If a human consciousness recognizes the android as a human, he/she will deal with it as a social partner even if he/she consciously recognizes it as a robot. At that time, the mechanical difference is not significant; and the android can naturally interact and attend to human society. Verification of this hypothesis is not easy and will take a long time. However, it is an important challenge that contributes to developing deeper research approaches in both robotics and cognitive science.[9]

One may wonder what is the vision driving the research in question. The "human-robot symbiotic society," what awaits us at the end of the experiment, is not a more *human* society, further humanized by the presence of robots, but the turning of humans into *inorganic*

---

[6] Hiroshi Ishiguro, "Android Science: Toward a New Cross-Interdisciplinary Framework," in *Robotics Research*, ed. S. Thrun, R. Brooks, and H. Durrant-Whyte (Berlin: Springer, 2007), 118.

[7] Ishiguro, "Studies on Interactive Humanoids," 72. The Total Turing Test (TTT) allows one to compare a robot manipulated by human operators and an autonomous robot controlled by developed technology.

[8] *Geminoid*, a tele-operated android, can transfer the presence of the person to distant places. Tellingly, Ishiguro suggests that "through tele-operation, the operator—myself—could adapt to the Geminoid body and accept it as my own body" ("Studies on Interactive Humanoids," 74).

[9] Ishiguro, "Android Science," 127.

intelligent life.[10] The evolution at stake is thus somewhat reversed with respect to a teleological movement geared toward the human being. The trajectory rather entails a progressive liberation of the latter from any "flesh body" and the openness, made possible by technology, to a diversity of bodily forms that "may allow us to evolve further." Evolution by technology pushes us beyond the limitations of life. "The ultimate aim of human evolution is immortality, achieved by replacing flesh and bones with inorganic material. Organic bodies are not a precondition for human existence in today's world….Humans come from, and return to, inorganic material. The human is currently almost a machine. We humans are going back to an inorganic state in the near future…we are trying to be an inorganic intelligent lifeform…. We can choose any kind of life forms…. That is, we can be released from the constraint of the human body."[11]

Can this world still be a world of persons? If so, what does it mean to retrieve a proper understanding of the person's singularity in a spectacle of world-objects? Is not the attempt to expand the notion of personhood to include robots a function of the totalizing tendency in which everything becomes an object? The levelling of the difference between person and machine would then signal the failure to account for the subject's separation from the world of objects, its *exteriority* to any totalizing pretense.

I take some of the notions elaborated by Levinas, whose echo is already evident in my *incipit*, as relevant to the question I am addressing in this paper. Mine will be less a direct confrontation with him, based on exegetical precision and punctual textual references, and more a personal appropriation of his mode of thinking, with which I find myself attuned.[12]

---

[10] Ishiguro, "Studies on Interactive Humanoids."

[11] Ishiguro, "Studies on Interactive Humanoids," 94–100.

[12] Levinas has been a "companion in thinking" for many years. I have dedicated my STL thesis to Paul Ricoeur ("Antropologia ed etica nella *philosophie de la volonté* di Paul Ricoeur," Pontificia Universita' Gregoriana, 1985) and my doctoral dissertation in theology to another phenomenologist, Dietrich von Hildebrand. See Roberto Dell'Oro, *Esperienza morale e persona: per una reinterpretazione dell'etica fenomenologica di Dietrich von Hildebrand* (Roma: Pontificia Universita' Gregoriana, 1996). Levinas has been an interlocutor for Ricoeur. On this, see Richard Cohen, *Ethics, Exegesis, and Philosophy: Interpretations after Levinas* (Cambridge: Cambridge University Press, 1992), 283–325. As for Dietrich von Hildebrand, the proximity to Levinas is evident at the level of their general sensibility, and this in spite of apparent reciprocal ignorance. Still, in a quasi-biographical narrative, concerning his encounter with phenomenology, and how he came to study with Edmund Husserl, Levinas mentions his friendship with Jean Héring, who was Husserl's student in Göttingen, together with "an entire circle of young thinkers." Since Hildebrand was part of it, Levinas must have known, at least, of the name. See Hans Reiner Sepp, ed., *Edmund Husserl und die Phänomenologische Bewegung. Zeugnisse in Text und Bild* (Freiburg/München: Karl Alber, 1988), 27–33. For a comparison between the two thinkers on their philosophy of love, see the introduction of John F. Crosby to Dietrich von

I begin with something like an archeological reconstruction of personhood in modernity, in order to locate the context out of which the question posed—"can a robot be a person?"—might take on meaning. Descartes, Hume, and Kant are the most important exponents of the story, their position emerging in direct contradiction to the classical metaphysics of the person, such as one finds in Thomas Aquinas. I see Levinas as having a complex relation with modernity, at once defined by a positive retrieval of Descartes and Kant, and critical of the anthropological dualism effected by the Cartesian *cogito*.[13] Levinas rejects the rationalist perspective of a *bodiless mind*, a person reduced to her cognitive capacities, no less than the empirical version of a *mindless body*, a person reduced to the external stimuli of sensations and impressions registered by the mind.

Contemporary bioethics, however, especially in the Anglo-American version of it, is mostly defined by such an understanding of personhood: the person does not "come to mindfulness" out of its bodily conditions, nor does she persist, when no longer conscious, in bodily presence. According to Peter Singer, one of the main voices in bioethics, personhood is transitional: it passes from being to being like a "thing," as long as certain dimensions of actual empirical consciousness are present. Thus, a dolphin, a chimpanzee, a higher mammal, perhaps a robot can be a person.[14]

On the other hand, as Levinas suggests, to be a person is to be "manifested in the exteriority of the face, which is not the disclosure of an impersonal Neuter, but *expression*, that is, the presence of an infinite idea that always exceeds the idea of the other in me."[15] If so, a robot cannot be a person. In what follows I try to say why I think this is the case.

Hildebrand's *The Nature of Love* (South Bend: St. Augustine's Press, 2009), xxxi. More broadly, the interesting article of Alexander Montes, "Toward the Name of the Other: A Hildebrandian Approach to Levinasian Alterity," *Questiones Disputatae* 10, no. 1 (2019): 82–109.

[13] More than on the various dualisms asserted by Descartes, Levinas focuses on the latter's idea of the infinite, as providing the point of entry into the meaning of the relation between same and other: "This relation of the same with the other, where the transcendence of the relation does not cut the bonds a relation implies, yet where these bonds do not unite the same and other into a Whole, is in fact fixed in the situation described by Descartes in which the 'I think' maintains with the Infinite it can nowise contain and from which it is separated a relation called 'idea of infinity'" (Emmanuel Levinas, *Totality and Infinity: An Essay on Exteriority*, trans. Alphonso Lingis [Pittsburgh: Duquesne University Press, 1969], 48).

[14] For a survey of various positions on personhood in bioethics, see the chapter on moral status in Tom L. Beauchamp and James F. Childress, *Principles of Biomedical Ethics*, 7th ed. (New York: Oxford University Press, 2013), 62–100.

[15] Levinas, *Totality and Infinity*, 50–51.

PERSON: A HISTORICAL RECONSTRUCTION

Facing the World in Wonder: Thomas Aquinas and the Person as Spirit Incarnate

For Aquinas, who comments on Aristotle's metaphysics, philosophy begins in *wonder*.[16] Wonder is the attitude that throws us back onto ourselves, in stunned astonishment at the sheer being there of things. Later on, such astonishment will engender perplexity about the meaning of things. Prior to the activity of questioning perplexity or doubt, wonder entails a kind of trust, a confidence (*fides*) in the natural goodness of reality, a confidence which is a love: being is promising and good (Genesis says: "God saw that it was very good"). For Thomas, who inherits the insights of the entire Christian metaphysical tradition, being is the miracle of gratuitous generosity, a "being there" without explanation, out of a source that gives. Creation is a gift of the Origin whose "coming into being" remains *in excess of* the ontological relations that define being in its "becoming," such as form and matter, formal and final causality, potentiality and actuality, etc. A being that is becoming already presupposes its "being given into existence," now charged with the promise of further development.[17]

But what does it mean "to be"? Thomas refers to two aspects of being: the "in-itself aspect" of being (*substance*), and its "towards-others aspect" (*relationality*). "To be" is "to exist," and to exist is to be *an-integrity-in-relation*. One might say that existence is the actualization of the energy of being, the gift from a source that offers itself, and whose communicative aspect is being participated to each existent.[18] Consider the following quotation from Gerald Phelan:

> The act of existence (*esse*) is not a state, it is an act, the act of all acts, and therefore must be understood as act and not as a static definable object of conception. *Esse* is dynamic impulse, energy, act—the first,

---

[16] The statement is originally from Plato, who lets Socrates say: "I see, my dear Theaetetus, that Theodorus had a true insight into your nature when he said that you were a philosopher; for wonder is the feeling of a philosopher, and *philosophy begins in wonder*," *Theaetetus* 155c-d, in *The Dialogues of Plato*, trans. B. Jowett (New York: Random House, 1937), vol. 1, 157. For Aristotle, "It is owing to their wonder that men both now begin and at first began to philosophize," *Metaphysics* 982b, in *The Basic Works of Aristotle*, trans. Richard McKeon (New York: Random House, 1941), 692. According to Thomas Aquinas, wonder (*admiratio*) is "a kind of desire (*desiderium*) for knowledge; a desire which comes to man when he sees an effect of which the cause either is unknown to him, or surpasses his knowledge or faculty of understanding" (*Summa theologiae* I-II, q. 32, a. 8).

[17] The third proof, so called "on possibility and necessity," speaks of the contingency of being. See Thomas Aquinas, *Summa theologiae*, I, q. 2, a. 3.

[18] "*Being* means that-which-has-existence-in-act…. Now any designated form is understood to exist actually only in virtue of the fact that it is held to *be*…. It is evident, therefore, that what I call *esse* is the actuality of all acts," in *An Introduction to the Metaphysics of St. Thomas Aquinas*, ed. James F. Anderson (South Bend: Regnery/Gateway, 1953), 22.

> the most persistent and enduring of all dynamisms, all energies, all acts. In all things on earth, the act of being (*esse*) is the consubstantial urge of nature, a restless, striving force, carrying each being (*ens*) forward, from within the depths of its own reality to its full self-achievement.[19]

Of course, one cannot fully grasp all this without reference to its theological underpinning. Thomas talks about being, but ultimately thinks about God, the Christian God of creation.[20] In this conceptual framework, the person also finds her place. Unlike other beings, the person is not just being-in-itself but being-coming-to-itself in self-presence (Thomas speaks of *reditio*, return unto itself).[21] The person is a mindfulness of being in its totality, a complete openness to its fullness and, thus, most fully being. Indeed, the person is the most perfect of substances, because of her ecstatic openness to the totality of being, an openness actualized in knowledge and will.[22] Such openness is not just a function of one particular aspect of the person, say the mind. It is the function of the entire being that is the person in the unity of its principle—i.e., its *soul* (for Thomas, like Aristotle: *anima est quodammodo omnia*)—because the soul is the form of the body (*anima forma corporis*).[23]

This is an important point: when thinking about the person, Thomas always points to the unity of body and soul. In encountering

---

[19] Gerald Phelan, "The Existentialism of St. Thomas," *Selected Papers* (Toronto: Pontifical Institute of Medieval Studies, 1967), 77. Quoted in Norris W. Clarke, *Person and Being* (Milwaukee: Marquette University Press, 1993), 9.

[20] To my knowledge, there is no reference to Thomas Aquinas in Levinas. The question of whether the former might not fall into Heidegger's general condemnation for the Western metaphysical tradition as "onto-theological" might have influenced Levinas' own way of reading the tradition. Jean-Luc Marion, a student of Levinas, is keen in rescuing Thomas Aquinas from the accusation. Consider the following: "To think *esse* starting from God, but not in inverse order (in the way of *metaphysica* and of Heidegger as well), allows Thomas Aquinas to free the divine *esse* from its—tangentially univocal—comprehension starting from what philosophy understands by being, entity being of the entity, in a word to mark the distance—an "infinitely infinite distance"—from the creature to God (Pascal)…. One could say that such (divine) *esse* keeps within itself the transcendence that opposes the act of being to the *esse commune* of entities' … Therefore, God without being (at least without *this* being) could become a Thomistic thesis." The quotations are from the essay titled "Thomas Aquinas and Onto-theo-logy," in Jean-Luc Marion, *The Essential Writings*, ed. Kevin Hart (New York: Fordham University Press, 2013), 306–307.

[21] "*Illa quae sunt perfectissima in entibus, ut substantiae intellectualis, redeunt ad essentiam suam reditione completa*," Thomas Aquinas, *De Veritate*, q. 1, a. 9.

[22] The point is central in the reinterpretation of Rahner and so-called "transcendental Thomism." See Karl Rahner, *Spirit in the World*, trans. William Dych (New York: Continuum, 1994).

[23] For a systematic analysis of Thomas Aquinas's anthropology, see the classic work of Sofia Vanni Rovighi, *L'antropologia filosofica di San Tommaso D'Aquino* (Milano: Vita e Pensiero, 1965).

the body, one encounters the person in her totality, in her spiritual presence. The person is, in this sense, "spirit incarnate," as Rahner would say, an integrity of being, a substantial unity. In this integrity of being the person is also infinite openness, infinite intentionality: *homo capax Dei*.

One might summarize by saying that for Thomas Aquinas the person entails three dimensions: to be an existent, that is, an integrity of being unto itself (incommunicable substance); actively open to other being (substance-in-relation); and passively receptive to the totality of being (receptivity). Incommunicability, relationality, and receptivity point to a balance between passivity and activity, *passio* and *conatus*, in the person.[24]

Here is where also the question of *potentiality* in the person comes into play. To be a person as substance-in-relation means that the person (like being itself) is always in a process of becoming: openness and receptivity are the conditions for growth and dynamic passing from potency to act. To be a person is always to be in-potency, in a process of progressive actualization toward a more perfect fulfillment. To be a person is to be an energy of transcendence.

The Thomistic synthesis will undergo a shift of paradigm as it moves into modernity, with important changes in the philosophical and theological disposition of the late Middle Ages, the age of nominalism. First, with respect to the concept of God: the stress is now on God's transcendence, on his distance from the world, as well as on his will, which is inscrutable and absolute—unlike Thomas, who sees the eternal law as the height of rationality. There are also changes in the concept of nature: both the distance of God and the inscrutability of his will leave the world devoid of signs of his presence. The world becomes "de-sacralized" or *secularized*, but, because of this, also available for the exploration of man, his engagement with things, and the general project of natural discovery. Nature can no longer hide—contra Heraclitus, for whom "nature loves to hide."[25]

God is beyond nature, infinitely powerful, and inaccessible. Now stripped of all the signs of divine communication, being is no longer conceived as communicative (*actus essendi*, act of existence), the energy of relationality, but as a being-unto-itself, distant and cold. A progressive impoverishment in the understanding of being sets in.[26]

---

[24] Clarke, *Person and Being*, 25–110.

[25] *Physis kruptesthai philei.* The fragment belongs to Heraclitus's *ipsissima verba* in the Diels-Kranz collection (B123). See Jonathan Barnes, *Early Greek Philosophy* (London: Penguin, 1987), 122.

[26] A thorough historical analysis can be found in Emerich Coreth, *Metaphysik: Eine Methodisch-Systematische Grundlegung* (Innsbruck: Tyrolia, 1980), especially 15–47. The impoverishment in the conception of being, following the nominalistic "effectual history," and in critical dialogue with Heidegger, has been argued by Gustav Siewerth, *Das Schicksal der Metaphysik von Thomas zu Heidegger* (Einsiedeln:

Consider as an example the loss of an analogical understanding of reality and the univocal reduction to particularity, taking place especially with Duns Scotus.[27]

One aspect of the change is paramount: the emphasis on *essence* rather than *existence*, and the subsequent loss of the notion of being as energy of communication. Being is no longer seen in the fullness of its over-determinacy, in its plenitude (Jacques Maritain speaks of the "generosity of being"), but as a thing-like presence, retracted unto itself, an object facing a subject. The essentialization in question signals the end of a teleological understanding of nature, the denial of its spontaneous finality. Nature knows no intrinsic destination. It is not a dynamic system that gives itself, as in the etymology of *physis*, but a static object, to be grasped by a subject.

THE PERSON AS SUBJECT AND THE DUALISM OF RENÉ DESCARTES

The story of modernity begins with a skeptical puzzlement about the nature of reality, the place of God in it, and the destination of the self. The modern *Denkform* is new with respect to its historical predecessors. If the classical tradition of metaphysics, from Plato to Aristotle and Aquinas, begins in wonder, Descartes's philosophical discovery, *cogito ergo sum*, emerges in the wake of doubt, and functions as the condition of possibility for the reconstruction of an image of the world now certain, finally resting on secure foundations beyond any skeptical assault.[28] There are losses and acquisitions following such a shift of paradigms.

The losses: Descartes no longer understands being in its character of act, energy, power, communicative to mind. Furthermore, being is no longer endowed with intrinsic value; as such, it was able to speak to the mind. Now a reversal of directionality in the relation between mind and being sets in: the mind dictates to being the conditions of its meaningfulness. "I think," that is, the epistemological certainty, precedes and grounds "I am," that is, the ontological constitution.[29]

Johannes, 1959). See also Etienne Gilson, *Being and Some Philosophers* (Toronto: Pontifical Institute of Medieval Studies, 1952).

[27] "Being is said in many senses" (*to on legetai pollakos*), says Aristotle in *Metaphysics*, Book IV, 2 (*The Basic Works of Aristotle*, 732). Similarly for Thomas, *ens multipliciter dicitur*.

[28] The reference is to the *Discourse on the Method*, Part IV, but the statement recurs in the *Second Meditation* as well. See *The Philosophical Works of Descartes*, trans. Elizabeth S. Haldane and G. R. T. Ross, vol. 1 (Cambridge: Cambridge University Press, 1981), 102.

[29] Marion has highlighted the ambivalence of Descartes's metaphysics, especially with respect to the idea of God. His focus is on Descartes's doctrine of "eternal truths," which cannot be ascribed both to the world and to God. Thus, all theology, understood as meaningful talk about God, is no more than a blank space or a white page in Descartes's writings. See Jean-Luc Marion, *Sur la théologie blanche de Descartes.*

I said there are also important theoretical acquisitions in this shift. The most important is the turn to subjectivity, an "anthropological turning point" which, later on, Kant will compare to a Copernican revolution: to be a person is to be a subject. In this revolution, the language of substance gives way to the language of subjectivity, self-reflexivity and interiority. The person is not just a substance, even if the most perfect, in ontological continuity with the others. It is an entirely different being, discontinuous with the rest of creation, because of its ability to think. Pascal will say: "Man is only a reed, the most feeble thing in nature, but a thinking reed."[30]

To the subjectification of substance follows the objectification of being: being becomes an object for a mind that is a subject. In its neutrality, being is simply a raw reserve of resources available for human exploitation.[31] Without a *telos*, an end (final cause), nature cannot account for its meaningful origin (*arche*). Nature becomes a neutral thereness, stripped of value, without formal or final cause.[32] It is the realm of effective causality, a network of forces linked together by mechanic interaction.

There is no denying that the main conquests of science rest upon such an understanding of reality. In his reconstruction of the entire trajectory of modern thought, especially in terms of its scientific advances, Edmund Husserl offers the following account:

> The exclusiveness with which the total world-view of modern man lets itself be determined by the positive sciences and be blinded by the "prosperity" they produced, meant an indifferent turning away from the questions which are decisive for genuine humanity. Fact-minded science excludes in principle precisely the questions which man finds the most burning: questions of the meaning or meaninglessness of the whole of human existence.[33]

In such a *Weltanschauung*, questions of meaning will, subsequently, be bracketed as unimportant—in fact, be expunged from the epistemic drive toward verifiability. What then remains for the

---

Analogie, création des vérités éternelles et fondement (Paris: Presses Universitaires de France, 1981).

[30] Blaise Pascal, *Pensées*, trans. A. J. Krailsheimer (London: Penguin, 1995), 347.

[31] William Desmond insists on such dialectic of subjectification and objectification in modernity. See his *Ethics and the Between* (Albany: SUNY Press, 2001), 17–47; and *God and the Between* (Oxford: Blackwell, 2008), 17–45.

[32] The modern gaze is akin to an act of unveiling, and the forcing of nature (Galileo) signals the end of teleology. So Spinoza, for whom "nature has no fixed goal and all final causes are but figments of human imagination" (*naturam finem nulllum sibi praefixum habere, et omnes causas finales nihil nisi humana esse figmenta*), Baruch Spinoza, *Ethics*, trans. Samuel Shirley (Indianapolis: Hackett, 1992), 59.

[33] Edmund Husserl, *The Crisis of European Sciences and Transcendental Phenomenology*, trans. David Carr (Evanston: Northwestern University Press, 1970), 5–6.

subject, when facing such a mechanized understanding of reality? It sees itself as other to the mechanical world or at least as irreducible to it. The world is purposeless, but the subject is purposeful, active with respect to being, not passive. It will provide being with the value being does not possess intrinsically. With Kant, the subject turns into "a self-assertive subjectivity." The person becomes the source of value, a noumenal being endowed with infinite value, with dignity.[34]

The separation of mind and being at the ontological level effects important changes in the understanding of the person. For the person too will now be defined by an intrinsic split, a separation between body and soul. The person is no longer grasped in the unity of a single substance but dissolved in the dualism of two separate substances. From the idea of the human being as a substantial unity of body and soul, as incarnate being, a separation sets in between soul (mind) and body. The soul and the body will be seen as separated from one another, with the soul losing its original meaning of principle of life and unity. Furthermore, the soul progressively becomes intellectualized as "mind," and this in opposition to the body.

The separation of mind and body opens up a dualism and a reduction in the understanding of the person: the person is her mind, independent of her body. The separation also comes with a new attribution of value: the mind is higher than the body. Here Descartes inherits the dualistic theological logic of modern devotionalism (*devotio moderna*), with its emphasis on the denial of the body.

What emerges is a kind of deracinated person, a personhood without body, a personhood reduced to its cognitive faculties. To be a person no longer means to grow into the biological space defined by a specific life-principle (the soul), but to be able to actualize the faculties that are proper to the traits of a "thinking substance." The person will be such only because *cogitans*, a capacity to think; and this independently of the body, whose connection with the mind is now viewed as entirely accidental. The human body, now separated from the spiritual principle, becomes like a machine (*res extensa*), which entails the loss of the incarnate self, and the intellectualization of the spirit. As the "other" substance in the human composite, the body is left open to the manipulative intervention of the superior principle. The emergence of anatomy in medicine focuses on the body as inert entity (*Körper*), rather than lived-reality (*Leib*).

We need to understand the profound implications of the dualism in question. We are inheritors of such dualism, and I will say that to ask about the personhood of a robot is to fall *de facto* into the trap set by such pre-comprehension. One can look at the consequences of the dualism in question from two angles: either the angle of a *bodiless mind* or the angle of a *mindless body*. We have here the two developments

[34] Desmond, *Ethics and the Between*, 17–47.

that follow Cartesian dualism: rationalism and empiricism. Let me begin with the latter.

THE DECONSTRUCTED SELF OF DAVID HUME

The empirical line, pursued by Hume and the British empirical tradition, develops all the way to nineteenth century utilitarianism. The body receives sensations, sensible impressions from experience and, progressively, the mind builds a sense of identity out of the congeries of such impressions.

What is the person? It is the product, the net result of external stimuli registered by the mind. One can see that because such impressions are seen in their *transitional* capacity to impress the mind, the identity of the person will only be the result of the *psychological*, rather than substantial, ability of the mind to retain impressions through memory. For Hume, the person is not an integral center of being (substance), but the flow of impressions that come and go, insofar as they are retained by memory. Thus, the paradox: there is "person" only insofar as there is actual consciousness of sensations and impressions or, as we might say today, empirical stimuli.[35]

Consider the thought experiments of contemporary bioethicists: when a person loses her consciousness, does she become a different person? Can we have two persons when a patient loses the ability to reason or to retain memories, etc.? The case becomes particularly acute when it comes to Alzheimer patients.[36]

For Hume, one of the great problems was the experience of sleep: when I am asleep, do I cease to be the person I was? Am I waking up every time a different person? Hume looked into the mind in order to find a unity to the flow of impressions and found nothing. He concluded that *there is no person* as a substantial principle of integrity-in-communication.

If there is a person only insofar as there is *actual* consciousness of sense impressions, then the body is no longer central in providing the condition for the mind's continuity, nor is the embodied-self understood as dynamically growing into its full potential. Consciousness of sense impressions means that both the preconditions of bodily development and the fading of mental functions in a still operative body will not affect the presence of personhood. The person does not "come

---

[35] In the *Treatise* (Book I, Part IV, Section VI), David Hume speaks of the self as "nothing but a bundle or collection of different perceptions, which succeed each other with an inconceivable rapidity, and are in perpetual flux and movement" (*A Treatise of Human Nature*, ed. L. A. Selby-Bigge [Oxford: Clarendon, 1988], 252).

[36] For an early example of the literature on this, see Allen Buchanan, "Advance Directives and the Personal Identity Problem," *Philosophy and Public Affairs* 17 (1988): 277–302. See also Helga Kuhse, "Some Reflections on the Problem of Advance Directives. Personhood and Personal Identity," *Kennedy Institute of Ethics Journal* 9, no. 4 (1999): 347–64.

to mindfulness" out of its bodily conditions, nor does she persist in its no longer conscious bodily presence. Prenatal life will not be personal. Nor will the life of the senile demented be the life of a person.

Two final points, one relative to freedom of the will, the other to the general conception of morality as based on emotions, rather than reason. For Hume there is no freedom of the will because the person, as he understands it, is entirely determined. Freedom of the will is a mask, a deception, at best the epiphenomenon of something else to which it can be reduced. Humean determinism will cast long shadows, all the way to contemporary philosophies of mind, in which the depersonalization of the human being is complete.

A second point: the deconstruction of personhood and the denial of freedom rest upon a conception of morality based on emotions, rather than reason. The feeling of sympathy towards other human beings will become the basis of morality. But how fleeting such a feeling is! The Humean retrieval of emotions, feelings and, in general, of the affective dimensions of the person, is *de facto* equivocal: it stands in the wake of Cartesian dualism and cannot provide a secure basis for moral judgment. Hume will become the father of non-cognitivism.[37]

THE PERSON AS UNIVERSAL EGO AND AUTONOMOUS SUBJECTIVITY: IMMANUEL KANT

I come to the rationalist line of development, following Cartesian dualism. Kant saw that the empirical ego is parasitical on a more original, *a priori*, notion of the self. The transcendental ego is the condition of all ordered experience *qua* experience of a unitary I.[38] Kant's philosophical anthropology represents a reaction to the fragmented self of Hume: the universal/transcendental ego is the formal capacity to gather the multiplicity, the manifold into an ordered unity.[39] With Kant, we have a more rigorous understanding of the meaning of personhood which, however, still stands within the "effectual history" of Cartesian dualism. Such dualism will be rendered by Kant in terms of a gulf between the phenomenal and the noumenal sphere of reality. To

---

[37] At the same time, the rehabilitation of the emotional sphere in Hume has not remained without important consequences, for example, in certain strands of contemporary feminist ethics. See Annette Baier, "Hume, the Women's Moral Theorist?," in *Women and Moral Theory*, ed. E. Feder Kittay and D. T. Meyers (Totowa, NJ: Rowman & Littlefield, 1987) and Alisa Carse, "The Voice of Care: Implications for Bioethical Education," *Journal of Medicine and Philosophy* 16, no. 1 (1991): 5–28.

[38] See Kant, *Critique of Pure Reason* on the "transcendental deduction."

[39] For the empiricist, the manifold in question can be brought to unity in a derived synthesis which, however, can easily fall apart without an *intrinsic* principle of unity. Such a principle points to a prior and not just derived synthesis, i.e., the "synthetic a priori": "The empiricist fails to see that a derived synthesis is possible only on condition of a prior synthesis, which is, in this respect, the condition of its possibility," William Desmond, *Desire, Dialectic, & Otherness: An Essay on Origins,* 2nd ed. (Eugene: Cascade, 2014), 69.

be a person is to belong to this noumenal realm, a realm of rationality and capacity for absoluteness. It is a realm of freedom, rather than natural necessity, such as the one entailed by Newtonian physics.

The noumenal realm also is a realm of value in a world stripped of it, as if the value sucked out of the objective world is now being channeled back into the subject, who becomes a being of infinite value. Question: where does the value of the subject come from if the world is valueless? How can something like a being *endowed with value* emerge from such a valueless world?

With Kant, the notion of personhood is clearly recognized in its moral significance. To be a person is to be a moral being able to exercise moral agency. In this resides the dignity of the person, its infinite value. To be a person is to be *autonomous*—i.e., to grant meaning, not to receive it (from God, religion, nature, etc.).

After Kant, the story of autonomy is the story of its progressive radicalization. There is a logical trajectory that runs from Kant to Nietzsche, from the person as power of autonomous decision making to the person as will to power.[40] In a world devoid of intrinsic value, and with an assertive (autonomous) notion of the person, the human also becomes an object. The objectification of the world turns into the objectification of the human. Having mechanized the world, we then end up mechanizing ourselves, looking at ourselves as mechanical systems, machines. The machine is the perfect expression of the modern outlook on reality: it is entirely constructed and available, that is, disposable. It is the perfect "object."

We too become like constructed mechanisms. We cry freedom of choice and autonomy but fall victim to all sorts of constructionisms: social (Marxism), psychological (Freud), neurological. Think of the attempt of cognitive science to reduce the human mind to the physiological circuitry of the brain, and explain away consciousness in terms of deterministic happening, entirely reducible to neurological functionality.

The essence of the machine: to be entirely passive to our own construction. In the case of the robot, its apparent activity, whether cognitive or practical, is entirely determined by us. Even when robots think for themselves, they do not really think. They only carry out preordained programs based on algorithmic laws of our own devise.

TO BE A PERSON: A PHILOSOPHICAL ACCOUNT

In light of this historical reconstruction, I now want to offer a more systematic understanding of the person. I begin with a provisional definition, something like a heuristic statement, aware of the fact that

[40] For this interpretation already, see Henri De Lubac, *Le drame de l'humanisme athée* (Paris: Cerf, 1999); and Romano Guardini, *Das Ende der Neuzeit. Ein Versuch zur Orientierung* (Würzburg: Echter, 1951).

Levinas would contest the possibility to "de-fine" a person. As a manifestation of the Infinite, the person escapes definition, at least in the sense that definitions are able to circumscribe objects. No objective definition of the person is possible without passing through a "lived subjectivity," the *event in being* Levinas calls "psychism": this is a *way of being* resistant to totality, in this case the one pursued by the search for definitional boundaries. Perhaps a better way to pose the question is not to enquire "what is a person?" but rather "who is a person?"

I think Levinas would push us to think of the person in terms of *an incarnate singularity, coming to itself, in openness to the Other*. I will parse out these three elements, incarnate singularity, selfhood, and openness to the Other, in order to arrive at an understanding of the person that owes greatly to Levinas, even if in the end it will differ from his in significant ways.

THE PERSON AS AN INCARNATE SINGULARITY

To be a person is to be individuated, to be an individual, to be one with oneself (already Boethius and Thomas Aquinas speak of *individua substantia*). Such individuality is expressive of a certain incommunicability (Roman law defines the person as *sui iuris et alteri incommunicabilis*). We are singular beings, though similar to others (brothers, monozygotic twins, or even clones). Levinas stresses this sense of incommunicability, independence, and separation, in terms of a break from any notion of participation:

> One can call atheism this separation so complete that the separated being maintains itself in existence all by itself, without participating in the Being from which it is separated—eventually capable of adhering to it by belief—One lives outside of God, at home with oneself; one is an I, an egoism. The soul, the dimension of the psychic, being an accomplishment of separation, is naturally atheist.[41]

The singularity in question is not the result of a statement of singularity; it is not a vindication of singularity (as "constituted," it would already belong to "the order of thought"). It is not constructed, but *given in the flesh*, in the body that we have, the "body that we are," to echo Gabriel Marcel. This singularity is incarnate. Only a human being can be a person—i.e., someone who belongs to the human species—because the singularity in question is embodied in the singularity of a human flesh.[42] Consider the following quotations from *Totality and Infinity*:

---

[41] Levinas, *Totality and Infinity*, 58.

[42] The question of the attribution of the notion of person to spiritual creatures, such as angels, or to God, would take us too far afield and cannot be addressed here. The focus remains on the human person. In critical distance from Heidegger, for whom

> The sensibility we are describing starting with enjoyment of the element does not belong to the order of thought but to that of sentiment, that is, the affectivity wherein the egoism of the I pulsates.[43]

> Sensibility established a relation with a pure quality without support, with the element. Sensibility is enjoyment. The sensitive being, the body, concretizes this way of being, which consists in finding a condition in what, in other respects, can appear as an object of thought, as simply constituted.[44]

This is relevant in addressing the question of double personality raised in the wake of the empiricist understanding. To say "I" (as we will see later, when talking about the self) is to actualize an energy of being that might still be dormant in the body, but that in its elementality reminds us of our being given to be *in the body that we are.*

To be a person is to be in a potential, constant state of growth. The person is always the promise of something more. We could say that such elementality cannot yet be objectified; it is pre-objective, in the sense of being felt, rather than determined in itself, as an object.

We feel ourselves, first of all, in bodily immediacy. If the dimension of self-insistence is prevalent at this point, it is a self-insistence that is also already a community with others. Levinas brings attention to the phenomenon of the *face*. In my face I say, "I am"; this always also is a "here I am," that is, "I am with." One must stress the flow-like, rather than fixed, character of this incarnate singularity.[45] The singularity in question is also a site of flow and passage, undergoing the world. The incarnate singularity is a passion of being.[46]

---

sensibility means the reversal of praxis over theory, Levinas stresses a more elemental notion of sensibility, which he calls "enjoyment" (*jouissance*): "Sensibility, Levinas discovers, does not first emerge as praxis caught up in the larger network of "in-order-to" (*das Um-zu*)—the "referential totality" (*Verweisungsganzheit*)—which ultimately implies Dasein. Rather, sensibility is first the sheer enjoyment of sensations, a 'care-free' contentment with sensing itself. Embodiment, sensibility, flesh, is, first a self-satisfaction and an enjoyment of elemental sensations, the sun on one's arms, the breeze in the air, indifferent to the higher-level significations of instrumentality and theory," Richard Cohen, *Ethics, Exegesis and Philosophy*, 154.

[43] Levinas, *Totality and Infinity*, 135.

[44] Levinas, *Totality and Infinity*, 136.

[45] Here is where a metaphysics of substance falters.

[46] One will note the difference between this elemental, suffering being (*passio*) and the self-positing ego of transcendental philosophy (*conatus*). Leibniz had a premonition of the meaning of this suffering being when he distinguishes between "perception" and "apperception": the self as flesh is perception not yet conscious in the distinct sense of apperception.

THE SELFHOOD OF THE HUMAN PERSON

The person is a being that comes to itself in self-awareness and reflectivity, in action and the development of moral agency. The self is not just given to itself. It becomes a self, it becomes a subject, indeed, a "thinking self" (Descartes). This coming to self, as Hegel points out, can only be possible in the intermediation with what is other. The third dimension in the definition provided above, that of "openness to the Other," is the last only in a temporal, rather than logical sense: it already subtends the other two components. The relation to the other is "older" than the relation to the self: "One may legitimately ask oneself whether the internal discourse of the *cogito* is not already a derivative mode of the conversation with the other; whether the linguistic symbolism that the soul uses in 'conversing with itself' does not suppose a dialogue with an interlocutor other than itself; whether the very interruption of the spontaneous impulse of thought reflecting upon itself, all the way down to the dialectical alterations of reasoning where my thought separates from and rejoins itself as if it were other than itself—whether this interruption does not bear witness to an *original and foregoing* dialogue."[47]

I said before that the incarnate singularity already undergoes the world. The body is always a medium of exchange, it is never only "mine": "The subjectivity of the subject, its very psyche, is a possibility of inspiration. It is the possibility of being the author of what has been breathed in unbeknownst to me, of having received, one knows not from where, that of which I am the author. In the responsibility for the other we are at the heart of the ambiguity of inspiration."[48]

FIRST ETHICAL SELVING: INTENTIONALITY

We come to ourselves in knowledge and action. The first component speaks to the exercise of mindfulness as an intentional act, a communion with what is other to us, an "object" toward which we move (or "to which we attend," *ad-tendere*) and yet could not do so, if not because of a mysterious participation already given to us in the ontological intimacy of mind and being.

This is why intelligence can never be "artificial": artificial intelligence is a preordained function wired to carry out certain operations. If we speak of intelligence, and can do so only analogically, we should always distinguish it from the intelligence of a person, a human being. Whereas a machine possesses, at best, "syntactical" capacity, the

---

[47] Emmanuel Levinas, *Of God Who Comes to Mind*, trans. Bettina Bergo (Stanford, CA: Stanford University Press, 1998), 146.

[48] Emmanuel Levinas, *Otherwise Than Being or Beyond Essence*, trans. Alphonso Lingis (Pittsburgh: Duquesne University Press, 1998), 148–49. "Inspiration" is existence "through the other and for the other, but without being alienation," (114–15).

person is capable of "semantic" appropriation—i.e., is able to understand what she is doing:

> In the sense in which people "process information" when they reflect, say, on problems in arithmetic or when they read and answer questions about stories, the programmed computer does not do "information processing." Rather, what it does is manipulate formal symbols. The fact that the programmer and the interpreter of the computer output use the symbols to stand for objects in the world is totally beyond the scope of the computer. *The computer, to repeat, has a syntax but no semantics*. Thus, if you type into the computer "2 plus 2 equals?" it will type out "4." But it has no idea that "4" means 4 or that it means anything else. And the point is not that it lacks some second-order information about the interpretation of its first-order symbols, but rather that its first-orders don't have any interpretations as far as the computer is concerned. All the computer has is more symbols.[49]

Like Searle, Hubert Dreyfus has been especially critical of the artificial intelligence model of human thinking and cognition understood as disembodied processes. Overall, his is a critique of dis-embodied artificial intelligence. Dreyfus uncovers a number of false presuppositions entailed by disembodied artificial intelligence. First, a biological assumption: the brain processes information in discrete operations by way of some biological equivalent of on/off switches. Second, a psychological assumption: the mind can be viewed as a device operating on bits of information according to formal rules. Third, an epistemological assumption: all knowledge can be formalized; what can be understood can be expressed in terms of logical relations. Fourth, an ontological assumption: since all information fed into digital computers must be in bits, the computer model of the mind presupposes that all relevant information about the world, everything essential to the production of intelligent behavior, must in principle be analyzable as a set of situation free determinate elements. His conclusion:

> Thus the view that the brain as a general purpose symbol manipulating device operates like a digital computer is an empirical hypothesis which has had its day. No arguments as to the possibility of artificial intelligence can be drawn from current empirical evidence concerning

[49] John R. Searle, "Minds, Brains, and Programs," in Boden, *The Philosophy of Artificial Intelligence*, 85 (emphasis mine). Searle's conclusion: "The point is that the brain's causal capacity to produce intentionality cannot consist in its instantiating a computer program, since for any program you like it is possible for something to instantiate that program and still not have any mental state. Whatever it is that the brain does to produce intentionality, it cannot consist in instantiating a program since no program, by itself, is sufficient for intentionality" ("Minds, Brains, and Programs," 87).

> the brain. In fact, the difference between the "strongly interactive" nature of brain organization and the non-interactive character of machine organization suggests that insofar as arguments from biology are relevant, the evidence is against the possibility of using digital computers to produce intelligence.[50]

Intelligence can only be an embodied act (think of the notion of "emotional intelligence"!), the actualization of knowledge and thinking through which we complete the world, we make the world come to itself, while receiving from the world the gift of a deeper sense of ourselves.[51]

The distinction between human and artificial intelligence becomes even clearer when we analyze basic dimensions of intentionality. Consider desire, imagination, and memory. In the eruption of *desire*, something possible only to an incarnate singularity, there emerges for the self the possibility to be other to itself. Desire reveals the energy of transcendence at work within the self and testifies to the power of self-differentiation in the self. There is more to the self than the identity of the same.[52] When the latter feels itself as lacking, it not only expresses something negative but, rather, more positively brings forth the energy of being in all its richness, exploding in the self in the form of desire. As Levinas says: "Desire is an aspiration that the Desirable animates; it originates from its 'object'; it is revelation—whereas need is a void of the Soul; it proceeds from the Subject."[53]

In this process, desire and *imagination* are allied. Imagination brings to further clarity the process of othering in the self, for desire acquires specific contours only in imagination. One ought to remember that the process of othering takes place in the intermediation with the other. The self is not fixed, it is a *metaphor*, a carrying across

---

[50] Dreyfus, *What Computers Still Can't Do*, 156.

[51] On this, consult the beautiful essay by Herman Krings, *Meditation des Denkens* (München: Kösel, 1956).

[52] For Paul Ricoeur to be a self is not to be the same, *idem* has to become an *ipse*! The issue of the relation between Ricoeur and Levinas is complicated; the two positions ought to be carefully distinguished. To put it briefly, and with reference to a pithy quotation from Ricoeur, the point of disagreement between the two consists in the fact that, for Ricoeur, "Awakening a responsible response to the other's call cannot work except by presupposing a capacity for reception, of discrimination, and of recognition" (*Oneself as Another*, trans. Kathleen Blamey [Chicago: The University of Chicago Press, 1992], 339). See Cohen's comment: "Following an existentialized version of philosophy's transcendental route, for Ricoeur, in contrast, there must always first be self-reflexivity, a capacity in the sense of a base, ground, zero-point, from which and out of which and into which otherness is *correlated*. For Levinas, in contrast, such an insistence on recognition, or on recognizing the priority of recognition, misses accounting for the prior impact which is at once the impact of alterity as such and moral obligation" (Cohen, *Ethics, Exegesis and Philosophy*, 304–05).

[53] Levinas, *Totality and Infinity*, 62.

which differentiates itself in the images of itself it images for itself. The self is the metaphor of a carnal mindfulness.

Careful vigilance is required, though, since the process of imagining itself as other to itself is equivocal. It can be seen as an end in itself, a process in which the self goes from one self-formation to the next, never encountering the other, and in so doing, never coming to itself. The self is then in flight from itself and not toward the other.

We end up dissipating the original energy of being, given to us in who we are, when we cannot face the selves we are in promise. *We need to remember* who we are in our self-transcendence. *Memory* is needed to balance desire and imagination, as well as the quest for self-transcendence: "Memory recaptures and reverses and suspends what is already accomplished in birth—in nature…By memory I ground myself after the event, retroactively: I assume today what in the absolute past of the origin had no subject to receive it and had therefore the weight of a fatality."[54]

Memory is the persistence of elemental self-awareness in the passage of transcending or becoming: "Memory as an inversion of historical time is the essence of interiority."[55] It is the return of the self to itself in the passage of becoming. I am talking here about memory in a non-objective way: not so much in terms of the process of remembering things, but as a kind of non-objective function, grounding our sense of interiority. Self-transcendence is not only externally directed: memory opens up the self to its inner otherness.[56] How is one to speak of artificial intelligence in terms of desire, imagination, and memory?

SECOND ETHICAL SELVING: ACTION

In action we come to ourselves as moral agents. This is first of all an openness, a response (*Wertantwort*) to the world and call of values which, for Levinas, is being revealed in the face of the Other, its vulnerability and indigence. To act morally is to transcend oneself; better, to embark in a movement of *transascendence*: "The metaphysical movement is transcendent, and transcendence, like desire and inadequation, is necessarily a transascendence."[57]

This in two ways: we transcend ourselves in what we become, when acting morally. This is why the "doing" involved in acting (*praxis*, *agere*) is different from the doing of production (*poiesis*, *facere*). In the latter, we do something that brings forth an external being, an external object. In the former, the effect is not outside the

[54] Levinas, *Totality and Infinity*, 56.
[55] Levinas, *Totality and Infinity*, 56.
[56] I am thinking of the discovery of the unconscious or Dostoevsky's "underground man" in *Notes from the Underground*: the ground of autonomy turns out to be a groundless abyss!
[57] Levinas, *Totality and Infinity*, 35. The term "transascendence," as Levinas clarifies in the footnote, is from Jean Wahl.

agent, but on the agent itself. The effect of moral action is the achieved integrity of agency.[58] The robot produces effects, but does not act in the sense above, thus it cannot be a moral agent.

The transcendence in question, however, is ultimately toward the other as other. For this reason, there can be an ambiguity at play here, when the process of moral performance turns into a journey of self-achievement, whereby the other is sought only as a function of one's fulfillment. We can desire the other out of a lack, now seen not so much as impelled by the energy of the source that originated us, but driven by the sense of lack that, negatively, determines the other for-self.

One might call this kind of transcendence "self-oriented." Hegel understands the subject in the process of its becoming thus, as self-determining negativity: the other is for-self, like in the master-slave dialectic of the *Phenomenology*, a story that repeats itself in Sartre's dialectic of masochism and sadism.[59] Deformation takes place with the stifling of the plenitude of excess, which is the origin prior to the lack of the self-oriented self, culminating with Nietzsche in the will to power willing itself for the purposes of its own self-glorification. Perhaps, with Levinas, one can envisage another way. For him "metaphysics does not coincide with negativity."[60]

THE PERSON AS OPENNESS TO THE OTHER

We need to understand the fulfillment of the self differently. While the self for sure is *penia* (poverty), it is also excess, because *porous* to the sourcing power.[61] The openness of transcendence can be *agapeic* (other-oriented), rather than self-oriented: it can be an openness for the sake of other-being, rather than self-being. The "agapeic self" breaks the circle of mediation and returns to the origin (God?) as a source of infinite energy. Impelled by the generosity of the origin, the self opens itself up to what is other to itself, breaching the circle of self-mediation. In self-oriented transcending, the self is more than the transcendence. It is *transascendence*: in other-oriented transcending, transcendence is more than the self.

A dialectical mediation is at play, one that becomes an inter-mediation, rather than a sublation of the other to the self (*à la* Hegel). The space between self and other rests intact as the middle space between

---

[58] The point, with the distinction between *poiesis* and *praxis*, is obviously Aristotelean.

[59] See G. W. F. Hegel, "Independence and Dependence of Self-Consciousness. Lordship and Bondage," in *The Phenomenology of Mind*, trans. J. B. Baillie (New York: Harper & Row, 1967), 228–40; Jean Paul Sartre, *Being and Nothingness: An Essay in Phenomenological Ontology*, trans. Sarah Richmond (London: Routledge, 2018), chap. 3: "Concrete Relations with the Other."

[60] Levinas, *Totality and Infinity*, 41.

[61] Think of the story of love in Plato's *Symposium*.

infinitudes: to the inward infinitude of the self, there corresponds the infinitude of the other. The mediation between such a plurality of infinitudes cannot be a self-mediation, in which one subordinates the other to itself. As the singularization of communicative being, the self *as self* cannot be reduced to will to power but will have to be understood as the *willingness to give itself* up for the other. Herein lies the paradox: loss of self is finding oneself. The "agapeic self" is dis-interested, in the sense of transcending self-interest into the middle (*interesse*). It is also *hetero-archic*: subject to the other, and because of this, a subject. In free obedience to the other, the self finally finds itself. For Levinas this openness to the other is the ultimate meaning of fecundity: "The I springs forth without returning, finds itself the self of an other: its pleasure, its pain is pleasure over the pleasure of the other or over his pain—though not through sympathy or compassion. Its future does not fall back upon the past it ought to renew; it remains an absolute future by virtue of this subjectivity which consists not in bearing representations or powers but in transcending absolutely in fecundity."[62]

The openness to the world, chiefly the world of the other, is an act of love, a fulfillment of reciprocity. To be a person is to love, because this is what "the incarnate singularity of a self, open to the Other" ultimately does: it actualizes itself beyond itself, in the responsibility for the other that is both a response (responsibility comes from *respondere*) and a release. A responsibility: I become myself, I become a moral agent, because I see myself commissioned by a call. "I am summoned as someone irreplaceable. I exist through the other and for the other, but without this being alienation."[63]

I can be many things, and yet fail myself, when failing to heed the particular call to which life calls me (this forms the nucleus of truth in situation ethics). Such call is singular: it is not a general call for "the humanity in me," as Kant would have it; nor is it a response reducible to the production of a good, not even "the greatest good for the greatest number" of utilitarianism: "Freedom is borne by the responsibility it could not shoulder, an elevation and inspiration without complacency. The for-the-other characteristic of the subject can be interpreted neither as a guilt complex (which presupposes an *initial* freedom), nor as a natural benevolence or divine 'instinct, nor as some love or some tendency to sacrifice.'"[64]

Whence the moral call then? Why should one heed it? Why be moral, in the end? Kant put the question in terms of a difference between a "hypothetical" and a "categorical" imperative. He had a premonition of the issue at stake here but was ambivalent about recognizing

---

[62] Levinas, *Totality and Infinity*, 271.
[63] Levinas, *Otherwise than Being or Beyond Essence,* 114.
[64] Levinas, *Otherwise than Being or Beyond Essence*, 124.

an alterity that summons the autonomy of the self, lest falling into heteronomy again. A communion of autonomous beings will be possible, for Kant, only in the "kingdom of ends." But this is only a postulate, an exigence of our thinking, not a reality we can know.[65]

The question here is the question of God as the ground of morality. An important question, and not only for Kant.[66] At stake is not only the question of God as the law-giver, who grants the moral imperative its absoluteness. It is rather a question of *release*: the release of our own freedom into the reciprocity of love, out of the love that generates us into being. We love because we are being loved. Generated into love, in the space of goodness predisposed for our enjoyment, we are capable of a freedom-for-the-other beyond autonomy, in the generosity of service. This is to be a moral being. This is, ultimately, what it means to be fully human, to be a person.

If I see correctly, Levinas has a tendency to moralize the relation to God and, subsequently, the notion of creation. He sees God as the infinite *in* the face of the other, but *not as the ground* of love for the same, which opens the same to the other. Because the other is the Master, "The interlocutor is not a Thou, he is a You: he reveals himself in his lordship. Thus, exteriority coincides with a mastery. My freedom is thus challenged by a Master who can invest it."[67]

There is moral earnestness in this God, but this is hardly a God of love! A different notion of creation is also needed. Levinas sees, correctly, that creation is a freeing of the person into her autonomy but does not quite understand such freeing against an ontology, a metaphysics of goodness. If this is the conclusion, then I am already *beyond* Levinas.[68]

---

[65] Immanuel Kant, "Transition from Popular Moral Philosophy to a Metaphysics of Morals," in *Grounding for the Metaphysics of Morals*, trans. James W. Ellington (Indianapolis: Hackett, 1981), 43.

[66] Recall Ivan's argument in Dostoevsky's *The Brothers Karamazov*: "If you were to destroy the belief in immortality, not only love but every living force that maintain the life of the world would at once be dried up…. For every individual…who does not believe in God or immortality, the moral law of nature must immediately be changed into the exact contrary of the former religious law." Fyodor Dostoevsky, *The Brothers Karamazov*, trans. Constance Garnett (New York: Barnes & Noble Classics, 2004), 71. The quotation is in Book 2, chapter 6 ["Why Is Such a Man Alive?"].

[67] Levinas, *Totality and Infinity*, 101.

[68] As William Desmond puts it, "Levinas seems to reiterate again and again, not only the horror of the *il y a*, but the *evil of being* in the relentless self-insistence of the *conatus essendi.* … His *version* of Plato's Good beyond being dictates a saving trauma and reversal from myself to the other. But *is there not an evil in that ethical good that sees being as evil?*…We have not quite been released to the agapeic 'It is good'" (*Art, Origins, Otherness: Between Philosophy and Art* [Albany: SUNY Press, 2003], 162). I have articulated the implications of such a metaphysics of the good for the field of bioethics in Roberto Dell'Oro, "On the Ultimate That Is the First: Thinking Beyond (Bio) ethics," *Gregorianum* 100, no. 3 (2019): 621–47.

CONCLUDING *UNSCIENTIFIC* POSTSCRIPT

At the end of this long detour on Levinas and the philosophy of the person, I return to the initial provocation and the question "can a robot be a person?" What to make, now, of the dream of an immaterial universe, the mad project instigated by a notion of robotics in which the human being becomes, in the end, "the subjugated subject"?[69]

In his 1920 novel *R.U.R.*, Czech writer Karel Capek describes the world imagined by scientist Rossum, a world of artificial workers (the word "robot" derives from the Slavic root for "work") intelligent and indefatigable, but incapable of feelings: "Rossum's Universal Robots," hence the acronym for the novel's title, "R.U.R." Although the intention of Rossum is to liberate human beings from the slavery of work and make them the masters of creation, the dream eventually fails: once all their needs are satisfied, human beings no longer have to work, and thus cease to reproduce. The robots, however, revolt and kill the humans, and their leader proclaims, "The time of man is over. A new world begins. The reign of robots."[70]

We are not there yet of course. And there is no need to think that robotics as a scientific enterprise has to necessarily end in transhumanist madness. Still, the retrieval of a personalist philosophy capable of highlighting the essential difference between person and machine provides an important buffer to any scientistic totalizing pretense. As long as this is the case, "the time of man is *not* over yet."

Levinas's phenomenological account provides important insights that remain closed to any scientistic gaze, including the one articulated by Ishiguro and his vision of a "human-robot symbiotic society."[71] For Levinas, as for Merleau-Ponty, phenomenology "is from the start a forswearing of science."[72] Thus the "unscientific" nature of this conclusion, which only calls for the suspension of those premises that make it impossible to see the person's singularity in a spectacle of world-objects. In this paper, I have attempted to provide something like an archeological reconstruction of such premises, only to show that the dualism they subtend falls short of accounting for what the person is: "an incarnate singularity, coming to itself, in openness to the Other." ℳ

Roberto Dell'Oro is the Austin and Ann O'Malley Chair in Bioethics, the Director of the Bioethics Institute at Loyola Marymount University, and professor in the Department of Theological Studies. He earned a doctorate in

[69] Thus Rémi Brague, in his reconstruction of modernity. See his *The Kingdom of Man: Genesis and Failure of the Modern Project*, trans. Paul Seaton (Notre Dame: University of Notre Dame Press, 2018), 160–68.
[70] See reference to Capek's novel in Brague, *The Kingdom of Man*, 167.
[71] Ishiguro, "Studies on Interactive Humanoids."
[72] Maurice Merleau-Ponty, *Phenomenology of Perception*, trans. Colin Smith (Oxford: Routledge, 1962), ix.

moral theology at the Pontifical Gregorian University in Rome and specialized in bioethics at Georgetown University. Roberto is the author/co-author of four books: *Pope Francis on the Joy of Love: Pastoral Reflections on* Amoris Laetitia (Mahwah, NJ: Paulist, 2018); *Health and Human Flourishing: Religion, Moral Anthropology, and Medicine* (Washington, DC: Georgetown University Press, 2006); *Esperienza morale e persona* (Rome: Gregorian University Press, 1996); and *History of Bioethics: International Perspectives* (San Francisco: International Scholars, 1996). He has translated Klaus Demmer, *Shaping the Moral Life* (Washington, DC: Georgetown University Press, 2000).

Journal of Moral Theology, Vol. 11, Special Issue 1 (2022): 157–181

Metaphysics, Meaning, and Morality: A Theological Reflection on AI[1]

Jordan Joseph Wales

ARTIFICIAL INTELLIGENCE is an increasingly pervasive, if hidden, factor in our daily lives. While "general" AI[2] remains, for the present, an aspiration rather than a reality, so-called "narrow" AI techniques now answer questions on our phones, translate between languages, select the advertisements that we see, recommend our next purchase or musical selection, identify potential hot-spots for crime, flag tumors in brain scans, and soon will drive us to work. Theology can and ought to say much about the ethical implications of artificial intelligence and our use of it, but I wish to ask: what may theology say about contemporary AI *in itself*? Some suggest that the answer is, "relatively little." Theologian David Bentley Hart contends:

> The operations of a computer are merely physical events devoid of meaning….[A] computer does not even really compute. *We* compute, using it as a tool….[I]ts operations are not determined by any semantic content but only by binary sequences that mean nothing in themselves. The visible figures that appear on the computer's screen are only the electronic traces of sets of binary correlates, and they serve as symbols

[1] For the development of this paper, I am indebted to too many persons to list, but I must thank the patience of the JMT editors, the insight of the two anonymous reviewers, as well as John Cavadini, Thomas Clemmons, Matthew Gaetano, Brian P. Green, Andrew Kuiper, Dwight Lindley, Anselm Ramelow, David C. Schindler, John Sehorn, Ezra Sullivan, Marga Vega, Marius Dorobantu, and John Seiffert. Each contributed important insights or commented on portions of the paper. The rest of you know who you are. The deficiencies of the final product have only me for their author.

[2] So-called "general" AI, the "ultimate goal of AI research," would be human-level or superhuman not in the sense of being conscious or having any sort of interior life—indeed, that is highly unlikely—but in being "applicable across all problem types." It would "[work] effectively for large and difficult instances while making very few assumptions." Needing "no problem-specific engineering," such a (now-hypothetical) system "can simply be asked to teach a molecular biology class or run a government. It would learn what it needs to learn from all the available resources, ask questions when necessary, and begin formulating and executing plans that work." Its success, then, would be in its behaviorally measured omnicompetence with respect to the goals that we appoint for it. See Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*, Reprint (New York: Penguin Books, 2020), 46.

> only when we represent them as such and assign them intelligible significances.[3]

It would seem, thus, that not only is it impossible for a programmed computer ever to constitute a mind of the sort that humans have; computers themselves are naught but systems of signs that exist *as* signs only at the whim of the beholder. Or, as Hart urges, they "have meanings only so long as they are objects of the representing mind's attention."[4]

Rhetorically, at least, this view is not without its difficulties. The claim that it is *we* who compute seems stretched to breaking by AI systems that convert Swedish into English or identify faces and fingerprints by self-generated formulae that even the systems' designers cannot comprehend. How can something have a merely observer-dependent meaning when it seems reliably tuned to the world in ways unfathomable to us? I argue that Hart's position—while true so far as it goes—is not so threatening to the reality or meaningfulness of computation as it may seem. As with printed text, we assign both functions and semantic content to tools and computers based on culturally shared intentional frames ("intentionality" here refers not to voluntariness but to "aboutness"). These framings determine the design (and our interpretation) of, for instance, a screwdriver's handle, a computer's output images, and the printed characters on a page. As observer-dependent realities, our artifacts are contingent, but they are not arbitrary.[5]

---

[3] David Bentley Hart, "Consciousness (Chit)," in *The Experience of God: Being, Consciousness, Bliss* (New Haven, CT: Yale University Press, 2013), 219.

[4] Hart, "Consciousness (Chit)," 218.

[5] Amie L. Thomasson, "Artifacts and Mind-Independence: Comments on Lynne Rudder Baker's 'The Shrinking Difference between Artifacts and Natural Objects,'" *APA Newsletter on Philosophy and Computers* 8, no. 1 (2008): 25–26. This is the case even for Piccinini's "mechanistic" account of computation, in which computation is defined not by any semantic content but by the manipulation of non-semantic machine states in line with some mechanistically specified rule. Semantic content can be assigned, of course, but it is not necessary to the definition of computation itself; see Gualtiero Piccinini, "Computers," *Pacific Philosophical Quarterly* 89, no. 1 (2008): 32–73, doi.org/10.1111/j.1468-0114.2008.00309.x; Gualtiero Piccinini, *Physical Computation: A Mechanistic Account* (New York: Oxford University Press, 2015). Even Piccinini, however, acknowledges that shared human purposes (in which we participate by our use of input/output devices) are necessary to fix the level of description wherein the mechanism is defined. Paul Schweizer therefore urges that even a mechanistic account is ultimately observer-dependent although not, therefore, arbitrary; see Paul Schweizer, "Computation in Physical Systems: A Normative Mapping Account," in *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence: Themes from IACAP 2016*, ed. Don Berkich and Matteo Vincenzo d'Alfonso (Cham: Springer International, 2019), 27–47, doi.org/10.1007/978-3-030-01800-9_2. Other positions might be taken, but at this time I find Piccinini, as modified by Schweizer, persuasive enough to move forward.

The fact of intentional framing indicates the ground upon which we may consider AI theologically. Observer-attributed meanings are tied up with the device's service to our purposes; an AI system mediates between our goals and the world with which it is engaged. Exploring these observer-dependent and world-attuned dimensions of AI in light of theological loci both moral (the spiritual life) and metaphysical (the doctrine of creation), I hope to facilitate further explorations of topics that, heretofore, have received comparatively little theological attention.

For this project I draw especially on Augustine of Hippo (lived 354–430 CE), who attended to human interpretation of the world within a Christian understanding of reality. His writings are respected by many Western Christian traditions and, on points relevant to my enterprise, are in broad agreement even with those Eastern traditions by which he is less esteemed. I make two claims:

First, with its metaphysics of *rationes seminales*, Augustine's theology of creation makes sense of the failures and successes of different AI methods by explaining the world as God's self-expression, a kaleidoscopic refraction of his Wisdom rather than a collection of discrete objects standing in crisp relations.

Second, these ontological considerations can be united to Augustine's account of interpretive judgment as a moral act bound up with love, in order to reveal the "deep neural network," contemporary AI's most powerful tool, as a kind of "memory" that maps the world to human purposes, without in itself accommodating the transcendent framing of the spiritual life. As such a "memory," the network may draw us to reduce reality to the measurable scope of this-worldly ambitions; or, as a *pointer* to reality, it may perhaps serve one's regathering of the created echoes of divine Wisdom as one journeys into the Trinity.

NATURAL WISDOM, OR AI'S CHALLENGE TO METAPHYSICS: WHAT IS THE WORLD?

"Symbolic" AI and its Ontological and Epistemological Failures

What computer scientists have called "artificial intelligence" has always reflected something of how their times have interpreted both human beings and the world. Somewhat following Thomas Hobbes, the dominant AI of the 1950s through the 1980s[6]—now called

[6] On the history of AI, see Nils J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements* (New York: Cambridge University Press, 2010). Or, popularly, see Luke Dormehl, *Thinking Machines: The Quest for Artificial Intelligence—and Where It's Taking Us Next* (New York: TarcherPerigee, 2017). The most widely used introductory textbook on AI is Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. (Upper Saddle River: Pearson, 2009). For clarity, I sometimes pass over distinctions that can be drawn between AI as human-like action (e.g., the Turing Test in the 50's), AI as human-like thought (e.g., Newell

"symbolic" or "Good Old-Fashioned" AI (GOFAI)—was philosophically founded on the "computationalist" hypothesis that thinking simply *is* the logical manipulation of symbolically represented information.[7] Accomplishing this task, a properly programmed computer would in fact *be* "thinking;" "a computer running a program that models a human cognitive process is itself engaged in that cognitive process."[8] Under this paradigm, a computer program that diagrammed a sentence and constructed a plausible response could be said to have *understood* that sentence.[9] Symbolic AI's greatest achievement was in "expert systems"—great structures of linked rules that, when queried, would generate a list of possible answers, perhaps posing further

---

and Simon's early work with symbolic representation in the 60's, leading to the field of cognitive modeling), AI as rational deliberation (e.g., logicism and expert systems in the 80's), and AI as rational agency (e.g., intelligent robots); on which see Russell and Norvig, *Artificial Intelligence*, 1–33.

[7] This intuition, a species of the Computational Theory of Mind (or "Computationalism"), was formalized as the "Physical Symbol Systems Hypothesis," seminally described in Allen Newell and Herbert A. Simon, "Computer Science as Empirical Inquiry: Symbols and Search," *Communications of the ACM* 19, no. 3 (March 1976): 113–26, doi.org/10.1145/360018.360022. The authors conclude: "Intelligence resides in physical symbol systems. This is computer science's most basic law of qualitative structure. Symbol systems are collections of patterns and processes, the latter being capable of producing, destroying and modifying the former. The most important properties of patterns is [sic] that they can designate objects, processes, or other patterns, and that, when they designate processes, they can be interpreted. Interpretation means carrying out the designated process. The two most significant classes of symbol systems with which we are acquainted are humans and computers" (Newell and Simon, "Computer Science as Empirical Inquiry," 125). For a recent assessment, see Nils J. Nilsson, "The Physical Symbol System Hypothesis: Status and Prospects," in *50 Years of Artificial Intelligence*, ed. Max Lungarella, Fumiya Iida, Josh Bongard, Rolf Pfeifer, vol. 4850 (Berlin: Springer, 2007), 9–17, doi.org/10.1007/978-3-540-77296-5_2.

[8] Jaegwon Kim, *Philosophy of Mind*, 3rd ed. (Boulder, CO: Routledge, 2010), 160. "Computationalism, or the computational theory of mind, is the view that cognition, human or otherwise, is information processing, and that information processing is computation over symbolic representations according to syntactic rules, rules that are sensitive only to the shapes of these representations. On this view …there is nothing more to a cognitive process than what is captured in a computer program successfully modeling it." Prominent advocates of some form of computationalism include Daniel Dennett and Steven Pinker; see Daniel C. Dennett, *From Bacteria to Bach and Back: The Evolution of Minds*, 1st ed. (New York: Norton, 2017); Steven Pinker, *How the Mind Works* (New York: Norton, 1997); and Steven Pinker, "So How Does the Mind Work?," *Mind & Language* 20, no. 1 (February 2005): 1–24.

[9] See Bertram Raphael, *SIR: A Computer Program for Semantic Information Retrieval*, AI Technical Reports (AITR-220) (Cambridge, MA: MIT, 1964), 42, hdl.handle.net/1721.1/6904. Even more comfortably asserting the identity of the computer's functioning with true understanding is Roger C. Schank and Robert P. Abelson, *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures* (Hillsdale, NJ: Erlbaum, 1977). This book occasioned John Searle's much-discussed rebuttal, the "Chinese Room" argument, in "Minds, Brains, and Programs," *The Behavioral and Brain Sciences* 3 (1980): 417–57.

questions to the user in order to prune the tree of possible resolutions. The most thrilling application of such systems was the Deep Blue chess computer that in 1997 defeated reigning world champion Gary Kasparov by winning two out of six games and playing to a draw in the other three.[10]

With time, however, symbolic AI came up against practical limits that suggested philosophical problems, particularly in the paradigm's underlying ontological and epistemological assumptions. Ontologically, symbolic AI worked with pre-defined sets of discrete categories standing in definite relations with one another. This diluted rationalism of innate ideas could easily implement Aristotelian syllogisms[11]—e.g., I wish to be dry in the rain; an umbrella will keep me dry in the rain; I will use my umbrella when it rains—but it did not yield a generalized capacity to deal directly with the world and human knowledge of it. Expert systems could break down in subtle situations wherein the interactions of tens of thousands of rules yielded unexpected and incorrect behaviors.[12] The incompletely understood congeries of factors bearing on the interpretation of a phrase or the outcome of an action made symbolic AI difficult to apply beyond constrained situations.[13] Nor could it represent or reason effectively about knowledge less precisely defined or more naturally contoured such as, for instance, one's sense of propriety in a social situation or one's route through a tangled wood.

Today, many problems of explosive scale in symbolic reasoning have been resolved or circumvented.[14] Guided by heuristic rules of

---

[10] On expert systems, see Nilsson, *Quest for Artificial Intelligence*, 229–40, 481–84. On Deep Blue, see Murray Campbell, A. Joseph Hoane, and Feng-hsiung Hsu, "Deep Blue," *Artificial Intelligence* 134, no. 1 (January 1, 2002): 57–83, doi.org/10.1016/S0004-3702(01)00129-1.

[11] See Aristotle, *Analytica Priora* (Selections), in *The Basic Works of Aristotle*, ed. Richard McKeon, Reprint (New York: Modern Library, 2001), I.2, 24b18–20.

[12] Nilsson, *Quest for Artificial Intelligence*, 326.

[13] Such problems embrace both "combinatorial explosion" (the intractable multiplication of factors in a rule-governed and search-based AI such as an expert systems) and the "qualification problem" (the impossibility of listing all preconditions for successful action). Combinatorial explosion was a special focus of the infamous Lighthill report, seen as responsible for a raft of funding cuts throughout Europe in the 1970s; see James Lighthill, "Artificial Intelligence: A General Survey," in *Artificial Intelligence: A Paper Symposium* (Science Research Council, 1973), www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm. These are related, in turn, to the "frame problem" (the impossibility of knowing *which* information is relevant and which can be ignored in the prediction of an action's effects).

[14] John McCarthy pioneered approaches to combinatorial explosion, the qualification problem, and the frame problem with his "Circumscription: A Form of Non-Monotonic Reasoning," *Artificial Intelligence*, Special Issue on Non-Monotonic Logic, 13, no. 1 (April 1980): 27–39, doi.org/10.1016/0004-3702(80)90011-9. Murray Shanahan wrote, recently: "Although improvements and extensions continue to be found, it is fair to say that the dust has settled, and that the frame problem, in its technical

thumb, logic engines like Doug Lenat's "Cyc" selectively traverse vast datasets of discrete categories and relations to analyze business practices and anticipate terrorist activity.[15] Still, the fundamental weaknesses of symbolic methods remain: They falter wherever discrete categories are difficult to detect, unknown, or too subtly intertwined. This is true for problems that humans handle poorly (e.g., weather prediction) and for those they solve well (e.g., behavioral prediction; language interpretation and translation; face recognition). Especially hard are those tasks in which humans attain to refined and effective sensibilities that, nonetheless, are difficult to articulate conceptually (e.g., aesthetics, improvisation, humor, and Go). In the words of Deep Blue architect Murray Campbell, human intelligence "is very pattern recognition-based and intuition-based," unlike symbolic AI's "search intensive" methods, which can require checking "billions of possibilities."[16]

Computer scientist and philosopher Brian Cantwell Smith argues that symbolic AI cannot provide a complete solution because its assumed ontology is inaccurate. The world, he writes, does not come "chopped up into neat, ontologically discrete objects" at human scale,

---

guise, is more-or-less solved" ("The Frame Problem," in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, Spring 2016 [Metaphysics Research Lab, Stanford University, 2016], plato.stanford.edu/archives/spr2016/entries/frame-problem/). The article cites Murray Shanahan, *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia* (Cambridge, MA: MIT Press, 1997); and Vladimir Lifschitz, "The Dramatic True Story of the Frame Default," *Journal of Philosophical Logic* 44, no. 2 (April 2015): 163–76, doi.org/10.1007/s10992-014-9332-8.

[15] Now deployed as Lucid.ai, developed by Cycorp Inc. See popular accounts in Cade Metz, "One Genius' Lonely Crusade to Teach a Computer Common Sense," *Wired*, March 24, 2016, www.wired.com/2016/03/doug-lenat-artificial-intelligence-common-sense-engine/; Doug Lenat, "Not Good as Gold: Today's AI's Are Dangerously Lacking in AU (Artificial Understanding)," *Forbes*, February 18, 2019, www.forbes.com/sites/cognitiveworld/2019/02/18/not-good-as-gold-todays-ais-are-dangerously-lacking-in-au-artificial-understanding/. For scholarly literature, see Douglas B. Lenat, "CYC: A Large-Scale Investment in Knowledge Infrastructure," *Communications of the ACM* 38, no. 11 (November 1, 1995): 33–38, doi.org/10.1145/219717.219745; Abhishek Sharma, Michael J. Witbrock, and Keith M. Goolsbey, "Controlling Search in Very Large Commonsense Knowledge Bases: A Machine Learning Approach," *Advances in Cognitive Systems* 4 (June 2016): 1–12; and Abhishek Sharma and Keith M. Goolsbey, "Simulation-Based Approach to Efficient Commonsense Reasoning in Very Large Knowledge Bases," *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 17, 2019): 1360–67, doi.org/10.1609/aaai.v33i01.33011360.

[16] Larry Greenemeier and Murray Campbell, "20 Years after Deep Blue: How AI Has Advanced since Conquering Chess," *Scientific American*, June 2, 2017, www.scientificamerican.com/article/20-years-after-deep-blue-how-ai-has-advanced-since-conquering-chess/.

"standing in unambiguous relations."[17] The world *seems* that way only because its ontological messiness has been made tractable by our human epistemology. Our ability to register the world, to apprehend it richly while coming to know individual objects, passing easily from sensation to conceptual thought, is something prior to the syllogism. Aristotle called this "abstraction."[18] In abstraction, something apprehended through the senses (e.g., this round taut-skinned tart-tasting misshapen sphere), comes to be understood consciously[19] as an instance of some more general category (e.g., plum)—that is, from sensation one comes to understand some *thing*. We do this easily, both recognizing objects and sensing their relations to one another, but it is ill accounted-for by the methods of symbolic AI, which proved clumsy and brittle when it came to distinguishing and identifying objects captured on camera or human words recorded through a microphone—tasks once expected to be easy in comparison to supposedly higher-level activities such as playing chess.

Crucially, according to Smith, our conceptualization of objects in the world is a form of *judgment*—not false but still deeply contingent and partial. We can meaningfully engage in discursive logical reasoning only because the abstractions flowing from our judgments remain grounded by our sense for their situatedness in a world not fungible with any finite set of symbols. More than a rule of thumb, this contextualization is necessary for true reasoning. Otherwise, as Smith says of symbolic AI, our conceptual symbolizations will "float free of

---

[17] Brian Cantwell Smith, *The Promise of Artificial Intelligence: Reckoning and Judgment* (Cambridge, MA: MIT Press, 2019), 28, 8. See also Smith, 34–35. I do not embrace all of Smith's metaphysical positions, but he is an exciting interlocutor.

[18] Aristotle, *De Anima*, in *The Basic Works of Aristotle*, III.4; *Metaphysica*, in *The Basic Works of Aristotle*, I.1; *Physica*, in *The Basic Works of Aristotle*, I.1. See also Allan Bäck, "The Conception of Abstraction," in *Aristotle's Theory of Abstraction*, New Synthese Historical Library (Cham: Springer International, 2014), 7–26, www.springer.com/us/book/9783319047584.

[19] "Consciously," i.e., as a conscious experience. While not agreeing with all of his positions, I will take philosopher John Searle's stab at a popularly accessible definition: "The central feature of consciousness is that for any conscious state there is something that it feels like to be in that state, some qualitative character to the state. For example, the qualitative character of drinking beer is different from that of listening to music or thinking about your income tax. This qualitative character is subjective in that it only exists as experienced by a human or animal subject. It has a subjective or first-person existence (or "ontology"), unlike mountains, molecules, and tectonic plates that have an objective or third-person existence. Furthermore, qualitative subjectivity always comes to us as part of a unified conscious field. At any moment you do not just experience the sound of the music and the taste of the beer, but you have both as part of a single, unified conscious field, a subjective awareness of the total conscious experience. So the feature we are trying to explain is qualitative, unified subjectivity" (John R. Searle, "Can Information Theory Explain Consciousness?," *New York Review of Books*, January 10, 2013, www.nybooks.com/articles/2013/01/10/can-information-theory-explain-consciousness/).

reality," potentially "devolv[ing]…into an endless play of signifiers, signifying nothing."[20] Not only do we need context; the world's richness and our sensitivity to it always exceed our explicitly stated concepts. We cannot, Smith argues, define a finite set of discrete categories—let alone define and detect in the real world the finite set of discrete features by which to identify something as belonging to those categories—that would lead to consistent and reliable performance for purely symbolic AI. There is more to the world, and more to thinking, than symbolic AI assumed.

"NON-SYMBOLIC" OR "STATISTICAL" AI

The problems cited above, along with immense advances in computing power, have brought recent eminence to so-called "non-symbolic" or "statistical" AI, a set of methods among which artificial neural networks hold greatest fame.[21] An artificial neural network is a computer program that mathematically simulates an interconnected set of simplified brain neurons. As an AI technique, then, it begins less from an interpretation of what human thinking *is* than from an analogy with its biological aspects. The goal of such networks is not so much human-like reasoning as it is neuron-like data-processing.[22] Having

[20] Smith, *Promise of Artificial Intelligence*, 73.

[21] The artificial neural network (ANN) was given its original form in Warren S. McCulloch and Walter Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity," *Bulletin of Mathematical Biophysics* 5 (1943): 115–33. For a time, this technique was neglected after Marvin Minsky and Seymour Papert's critique of single-layer networks' inability to perform certain elementary logical functions (e.g., XOR); see *Perceptrons: An Introduction to Computational Geometry*, 1st ed. (Cambridge, MA: MIT Press, 1969). Fifteen years later, a method for training multi-layer networks was described in David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, "Learning Representations by Back-Propagating Errors," *Nature* 323 (October 1986): 533–36, doi.org/10.1038/323533a0. Nonetheless, Minsky and Papert released an "Expanded Edition" of their book in 1987, refining and restating the limitations of ANNs. *Perceptrons* is often accorded a causal role in the "AI winter" of the 70s through the 90s, a decline of research in light of the perceived limits of both "symbolic" methods and ANNs; see Mikel Olazaran, "A Sociological Study of the Official History of the Perceptrons Controversy," *Social Studies of Science* 26, no. 3 (1996): 611–59, www.jstor.org/stable/285702. The current renaissance of ANN techniques, specifically "Deep Learning" (neural networks with many layers) began in 2012 with AlexNet, a deep convolutional network capable of amazing feats of image recognition; Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM* 60, no. 6 (May 24, 2017): 84–90, doi.org/10.1145/3065386; Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning," *Nature* 521 (May 28, 2015), www.cs.toronto.edu/~hinton/absps/NatureDeepReview.pdf.

[22] To say that ANNs are non-symbolic does not mean they are irreconcilable with computationalism. Moreover, perhaps they could even be considered "symbolic" at the appropriate scale. See discussion of these two controverted issues in Michael Rescorla, "The Computational Theory of Mind," in *The Stanford Encyclopedia of Philosophy*, plato.stanford.edu/archives/spr2017/entries/computational-mind/.

some affinity with the British empiricist tradition, these methods are far less beholden to assumptions about either ontology or epistemology than are the techniques of symbolic AI.[23]

An artificial neural network receives a pattern of information as numerical values at its input nodes, which are connected with various strengths to layer upon layer of further nodes. At each node, when the sum of incoming connections exceeds some pre-set threshold, that node will fire and its own signal will be transmitted variously to nodes on a further layer, and so on. If you put in a pattern at the beginning, it is transformed as its elements are recombined and processed until something else comes out on the final layer of the network. A network can be "trained" to produce desired responses—say, to predict travel patterns or to recognize faces—by adjusting the strengths of its connections, thus tuning the contribution made by each node to each recombination and, in due course, to the final result. A piano offers a poor analogy but a useful image. If you have ever shouted into the instrument with its sustaining pedal held down, then you have heard its tuned strings resonate with the different frequencies of your shout. One receives back a sort of echo, not of one's words but of the tones of one's voice. Similarly, as a neural network is tuned (i.e., as its connection strengths are adjusted), it begins to resonate with the entangled relations implicit in our world, including relations not easily discerned or logically represented by human investigators. Moreover, by its training, the network does not just echo; it transforms input data in order to make explicit the relations that are of interest to the trainer.

Neural networks and other statistical methods subserve the AI that underlies self-driving cars, programs that beat world champions in the games of Go and chess,[24] the voice recognition of Siri and Alexa,[25]

---

[23] They are not wholly empiricist, but have certain predetermined architectural features, with the debate centering on whether these are domain-general features (as empiricists would claim to be the case in the human brain) or domain-specific, which would entail some "nativist" or quasi-rationalist innateness in their "interpretive" action; thus Cameron Buckner, "Deep Learning: A Philosophical Introduction," *Philosophy Compass* 14, no. 10 (2019): 11–12, doi.org/10.1111/phc3.12625.

[24] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis, "Mastering the Game of Go without Human Knowledge," *Nature* 550, no. 7676 (October 19, 2017): 354–59, doi.org/10.1038/nature24270; David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis, "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm," *ArXiv:1712.01815 [Cs]*, December 5, 2017, arxiv.org/abs/1712.01815.

[25] Sree Hari Krishnan Parthasarathi and Nikko Strom, "Lessons from Building Acoustic Models with a Million Hours of Speech," *ArXiv*, no. 1904.01624 (Cs, Eess, Stat), April 2, 2019, arxiv.org/abs/1904.01624; Brian Barrett, "Alexa's Had a Big Year,

Google Translate,[26] webmail autocomplete functions,[27] and the "curated" recommendations delivered by Spotify, Netflix, and Amazon.[28] Many problems that bedevil symbolic methods can be solved handily by a neural network because, in a manner of speaking, the network is receptive to, imprinted by the structure of the world as presented to it. We might say that it develops a point of view: not a conscious experience, but something like the classical notion of the mind's conformity to a thing[29]—although here that conformity is always constrained by the task for which the AI is trained. But *to what* is it conformed? To answer that question, we need a richer ontology.

CONCEPT AND CONTEXT

Symbolic AI's treatment of the world has a long pedigree that finds analogues in certain streams of ancient Greek thought, for which to

---

Mostly Thanks to Machine Learning," *Wired*, December 19, 2018, www.wired.com/story/amazon-alexa-2018-machine-learning/.

[26] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," *ArXiv*, no. 1609.08144v2, September 26, 2016, arxiv.org/abs/1609.08144; Cade Metz, "An Infusion of AI Makes Google Translate More Powerful Than Ever," *Wired*, September 27, 2016, www.wired.com/2016/09/google-claims-ai-breakthrough-machine-translation/; Gideon Lewis-Kraus, "The Great A.I. Awakening," *The New York Times Magazine*, December 14, 2016, www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html; Douglas Hofstadter, "The Shallowness of Google Translate," *The Atlantic*, January 30, 2018, www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/.

[27] Yonghui Wu, "Smart Compose: Using Neural Networks to Help Write Emails," *Google AI Blog* (blog), May 16, 2018, ai.googleblog.com/2018/05/smart-compose-using-neural-networks-to.html.

[28] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah, "Wide & Deep Learning for Recommender Systems," in *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems - DLRS 2016* (Boston: ACM, 2016), 7–10, doi.org/10.1145/2988450.2988454; Faisal Siddiqi, "Machine Learning Platform Meetup: Recap of the Oct 2017 ML Platform Meetup at Netflix HQ," *Netflix TechBlog* (blog), October 18, 2017, medium.com/netflix-techblog/machine-learning-platform-meetup-ddec090f3c17.

[29] E.g., Thomas Aquinas, ST I, q. 16, a. 1, co.: "Knowledge is according as the thing known is in the knower" and the "truth [of one's own thoughts] is the equation of thought and thing." See also ST I, q. 16, a. 3; translation from *Truth: A Translation of Quaestiones Disputatae De Veritate*, trans. Robert W. Mulligan (Chicago: Regnery, 1952), 1.1. One's apprehension of the world is not just a symbolic representation of an account of it but is a world-conformed habit of mind from which such accounts and their representations are generated. One's capacity for understanding is shaped by one's experience and one's memory and accompanies one in every experience.

understand a thing—be it a natural object or a human-made artifact—was to apprehend rationally its "form" or "idea," that is, its configuration toward some activity or use. According to Langdon Gilkey, for such thinkers,

> Once the scientist has uncovered this form of the object….he really knows all he can or need know about it; he has penetrated to the very heart of reality. He does not need to experience or describe any further its external characteristics or patterns of behavior, since, with its form in his mind, he can predict all that is important about its activities and powers.[30]

Indeed, all "sensible characteristics" (e.g., a knife's gleam) beyond those necessitated by the form (e.g., its cutting edge) are but byproducts of the "necessary but distorting…substratum [that has been] arranged according to the guiding principle" of the form.[31] They may be discarded from consideration as meaningless "result[s] of unpredictable flaws in the material and so quite beyond rational explanation."[32]

This (perhaps unnuanced) rendering of ancient Greek science strikingly anticipates the formalistic world-model assumed by symbolic AI, which Smith finds inadequate both to physical realities and to how we apprehend them: "Taking the world to consist of discrete intelligible mesoscale objects is an *achievement* of intelligence, not a premise on top of which intelligence runs."[33] The concepts by which the world is discretized (i.e., Gilkey's forms) *do* represent reality but they are engagements *with* it rather than separable simulations of it. They are instruments by which we interact with the world at a particular but non-exhaustive level of description. Only as points of contact with their real-world context do they remain true to it. Therefore, especially in "long chains of articulated reasoning" about realities unavailable to immediate experience, our highly abstracted formal concepts must remain habitually "embedded" in their underlying "sub-conceptual webs" so that, from this context, we may draw the "subtleties, adjustments, and so on" that will give "nuance and inflection" to inferences both immediate and distant.[34] By this embedding, articulated reasoning can be a true engagement with rather than a reduction of the world.

Smith derives his understanding of the "sub-conceptual" from the success of today's "deep" neural networks, which have broad input layers and dozens of interior layers. "When fed with data obtained

---

[30] Langdon Gilkey, *Maker of Heaven and Earth: A Study of the Christian Doctrine of Creation* (New York: Doubleday, 1965), 124.
[31] Gilkey, 123–24.
[32] Gilkey, 126–27.
[33] Smith, *Promise of Artificial Intelligence*, 35.
[34] Smith, *Promise of Artificial Intelligence*, 74–75. On these sub-conceptual webs, see also Smith, *Promise of Artificial Intelligence*, 34–35.

directly from low-level sensors" such as cameras, the "high-dimensionality" of the network's layers enables it "to 'encode' all kinds of subtlety and nuance …[without] hav[ing] to categorize and discretize their inputs at the outset." Its self-adjusting weightings, which "store and work with" the input, come subtly to reflect relations between phenomena in the data. Like the cultivation of a sommelier's palate, a progressive attunement to the world's "'sub-conceptual' terrain" renders the network effective in a way that pre-categorization by formal ontologies would never have permitted.[35]

The nature of this attunement, however, is difficult to explore because, as physicist Judea Pearl writes, neural networks are "opaque." Even when tuned to their training data and able to generalize to new data, their interior sensitivities are not at all easily interpreted.[36] Eminent computer scientist Peter Norvig argues that their statistical attunement "describes what *does* happen" but—"mak[ing] no claim to correspond to the generative process used by nature"—it "doesn't answer the question of *why*."[37] Norvig's statement is of ambiguous value. True, networks do not develop theoretical models, but if Smith is right, then the network may say quite a bit, even if obscurely. How could nature's "sub-conceptual" *not* be somehow related to the deep flow of its "generative process[es]," especially if this sub-conceptual gives *our* theoretical concepts their success as engagements with nature itself?

To develop a joint account of nature and networks, let us reflect on what the "sub-conceptual" might be. Consider a hypothetical (but technologically realistic) neural network, trained to distinguish grasses, wildflowers, and trees with fidelity to distinct scientific categories.[38] In "earlier" layers, we might observe activity quite out of keeping with these hierarchical classifications as, for instance, if certain areas were to be equally highly activated by the subtle ridging on a blade of grass, the stalk of a valerian wildflower, and the needles of certain conifers. Were this but a matter of surface-level similarity with no more conceptual heft than the redness of a coral snake grouped with that of a red panda, then we might agree that the network's activity "bear[s] no relation" to nature's "generative process[es]." But what if

---

[35] Smith, *Promise of Artificial Intelligence*, 58–59. It has been proposed that deep networks find inherent symmetries in the data manifolds, to yield useful relations and to encode a large amount of this data. See Buckner, "Deep Learning," 9–11.

[36] Judea Pearl, "The Limitations of Opaque Learning Machines," in *Possible Minds: Twenty-Five Ways of Looking at AI*, ed. John Brockman, 1st ed. (New York: Penguin, 2019), 18.

[37] Peter Norvig, "On Chomsky and the Two Cultures of Statistical Learning," 2011, norvig.com/chomsky.html. Emphasis original.

[38] As critics rightly point out, at no point does the network learn to see these plants as wholes; see Gary Smith, *The AI Delusion* (New York: Oxford University Press, 2018), 50–51. Still, my point concerns that to which it is attuned within wholes.

the network has captured features that give us access to Smith's "sub-conceptual web"? The visual similarities of grass, stalk, and needle are not mere surface coincidence. A ridged configuration strengthens these plants' narrow structures, which are mostly hollow but must remain stiff to perform their function. In grouping these three, then, the neural network is attuned to what *we* know to be a manifested harmonization of gravity, force of wind, capillary action, and—in the movement of fluid—a hint of metabolism. These three plants are not at all closely related nor all common to the same environment; yet, as expressions of this kingdom of life, they have settled into a groove that expresses something not only about these particular organisms or even about their local environment, but also about the natural harmony of earth as a whole.

All this is taken into the absorbent mind of the attentive child; it is forgotten amidst the classifications by seed, climate, and species in an introductory biology class; and it is rediscovered by the botanist and the gardener. It is like the sounding of a piano note, which bears witness not only to the struck key but also, in its overtones, to the shape of the piano, the species of wood from which it is constructed, and even—perhaps discernible only by a neural network—the orientation of the grain and the history that imparted to that particular tree its distinctive physical quirks.

For a reductively formalistic science, the three plants' ridges mislead because we would define their forms better by macroscale physical characteristics, climates, and modes of nutrition and reproduction. For the same science, *all* that I have written of the piano falls among the "sensible characteristics" that may be discarded upon grasping the intelligible form of the keyed instrument. Against such a view, I contend that *this* sort of thing is what we have meant by "piano" all along; and it is why the classically trained pianist finds something lacking in the finest electronic instrument, as a matter not of snobbery nor of tradition only, but of the full meaning of the piano's "form." Like the commonalities of the three plants, the distinctiveness of the instrument *cannot* be captured by abstracting from its sensible characteristics because its *truest* concept—the concept that we hold—is adequately transmitted only by the experience of the piano itself as a transduction of the world from which it is drawn. It is not that our formal concepts fall short of experience; it is that, as we see in the sensitivity of the neural network, they embrace much more of reality than our way of speaking may lead us to believe.

The Rationes Seminales, an Augustinian Ontology Commensurate with AI

If our attempts to "purify" the conceptual from its sensuous matrix lead to a parody rather than to a more precise grasp of reality, and if concepts engage a thing's form, then I propose to think of particular

forms as including, rather than abstracting from, the harmonics of nature. Augustine of Hippo has in mind such an inherently dynamic form when he speaks of a thing's *ratio* (in Greek, *logos*). His ontology of *rationes* (*Gen. ad litt*. 4–6)[39] makes sense both of symbolic AI's failure and of neural networks' successes by describing how particular things are inextricably at home in the world. Meanwhile, as I will discuss later, his teaching on interpretive judgment (*Trin*. 9.6.11–9.11.16, 15.10.17–15.11.21) clarifies how concepts are "achievements"[40] that are truest engagements when they do not detach things from that matrix.

The key for Augustine is contingency. Plato and the tradition emanating from him sought a fixed non-contingent world—i.e., the forms or ideas—in light of which the contingent and the shifting might be explained. For Augustine, however, there is no world of the forms. The only non-contingent reality is God himself, a simple being, alive in love. Construed as the archetype of all things, his transcendent and inexhaustible life is the divine Wisdom (Prov 8), in which are the non-contingent prototypes or "eternal reasons" (*aeternas rationes*) of all contingent things—not as distinct forms but as identical with his simple life (*Trin.* 12.2.2). "God would not make creatures unless he knew them before he made them; nor would he know them unless he saw them; nor would he see them unless he possessed them; nor would he possess what had not yet been made except as uncreated being, as he is himself" (*Gen. ad litt.* 5.16.34).[41] God's uncreated life is single and simple. Therefore, the *aeternas rationes* are distinguishable only from our point of view, being aspects of God seen as "simply multiple and uniformly multiform" through the prisms of his contingent *created* expressions here below (*Ciu.* 12.19).[42]

---

[39] See, among the secondary literature, Gerald P. Boersma, "The *Rationes Seminales* in Augustine's Theology of Creation," *Nova et Vetera* 18, no. 2 (2020): 413–41, doi.org/10.1353/nov.2020.0030; Christina Hoenig, "Augustine," in *Plato's Timaeus and the Latin Tradition* (Cambridge: Cambridge University Press, 2018), 242–51; Luigi Gioia, *The Theological Epistemology of Augustine's* De Trinitate (New York: Oxford University Press, 2016), 262–69. Also, setting the *rationes* in wider context for today's inquiries, see John C. Cavadini, "Augustine and Science," in *T&T Clark Handbook of Christian Theology and the Modern Sciences*, ed. John P. Slattery (London: T. & T. Clark, 2020), 59–66.

[40] Already quoted from Smith, *Promise of Artificial Intelligence*, 35. See this paper, note 33.

[41] Augustine of Hippo, *The Literal Meaning of Genesis* (401–415), trans. John Hammond Taylor, vol. 1, ACW 41 (New York: Paulist, 1982), 167.

[42] Augustine of Hippo, *Concerning the City of God Against the Pagans* (413-427), trans. Henry Scowcroft Bettenson, Penguin Classics (London: Penguin Books, 2003). See also *Gen. ad litt.* 5.13.29–5.15.33, especially 5.15.33, translated in Augustine, *Literal Meaning of Genesis*, 1:166: "What has been made through Him is understood to be 'life' in Him, the life in which He sees all things when He makes them. He has made them as He has seen them, not looking beyond Himself, but He has numbered within Himself all that He has made. His vision and that of the Father are not different:

In the contingent world, divine Wisdom is expressed doubly: in the kinds of things created according to the *rationes*; and in God's providential governance of the whole, whereby the ways of these things are expressed in interaction with one another (*Gen. ad litt.* 5.12.28).[43] As for kinds, the nature and capacities displayed in the life of each created thing reflect, facet-like, the goodness and wisdom of God:

> Through Wisdom, all things were made; and the motion we now see in creatures, measured by the lapse of time as each one fulfills its proper function, comes to creatures from causal reasons [*rationes*] implanted in them, which God scattered as seeds at the moment of creation when *He spoke and they were made; he commanded and they were created* [Ps. 32:9] (*Gen. ad litt.* 4.33.51).[44]

While distinguishable, created kinds are not isolatable. Somewhat as all created *rationes* are found archetypically in the one divine Wisdom, each plant and animal has its common origin in the causality of the one earth that "received the power of bringing them forth" (*Gen. ad litt.* 5.4.11, see Gen 1:12).[45] From the very beginning, the universe has contained, *in nuce*, the meaningfulness that historically has unfolded into the distinction of contingent creatures. Thus, God made "all things together" (Sirach 18:1; *Gen. ad litt.* 5.23.44–46).[46]

The second contingency—which Augustine honors as no purely platonic thinker could—is history itself.[47] This is the sphere of God's

---

there is one vision, as there is one substance"; citing Job 28:12–13, 22–25. See also, plainly showing that these are not "moments" in God's life, but the eternal life that is God's existence, *Trin.* 4.3. Augustine affirms that nothing is "irregular or unforeseen" by God, because the "*rationes* for all things created and about to be created are contained in the mind of God," "eternal and…unchangeable;" *Ciu.* 12.19. See also John C. Cavadini, "God's Eternal Knowledge According to St. Augustine," in *Cambridge Companion to Augustine*, ed. David Vincent Meconi and Eleonore Stump, 2nd ed. (New York: Cambridge University Press, 2014), 37–59.

[43] Augustine distinguishes the unchangeable *rationes* from God's work from which he rested (i.e., creatures, with their immanent *rationes*) and the things he produces from these works—that is, material things and their motions under providence according to their particular *rationes*.

[44] Cited by Cavadini, "Augustine and Science," 64.

[45] Augustine, *Literal Meaning of Genesis*, 1:153. See also *Gen. ad litt.* 5.4.11; 6.6.10–11; 6.10.17; 6.14.25; Ernan McMullin, "Evolution as a Christian Theme" (Herbert Reynolds Lecture Series, Baylor University, 2004), 7–8.

[46] Augustine does not seem to think that the distinct *rationes* are contingent *within* our historical frame. In other words, while his theory is ripe for development into a theology of biological evolution, he himself does not fully anticipate it.

[47] The Christian belief in God's progressive self-revelation culminating in the Incarnation would have sensitized Augustine to history. We find this even in his early and supposedly neoplatonic treatise *De uera religione*; on which see recently Thomas Clemmons, "The Common, History, and the Whole: Guiding Themes in *De Vera Religione*," *Augustinianum* 58, no. 1 (June 28, 2018): 125–54, doi.org/10.5840/agstm20185816.

providential governance, not a marionette-like foisting of the divine will upon otherwise-free creatures, but an elicitation of their interacting harmonies by the unfolding of temporal events (*Gen. ad litt.* 5.11.17; 5.20.41; *Trin.* 3.5–6.11).[48] Not only does the general developmental and behavioral history of squirrels manifest more fully their *ratio* as a refraction of divine Wisdom; Wisdom is further manifested through particular things' contingent histories of interaction—e.g., *this* squirrel in *this* forest, scrambling up *this* oak tree away from *that* fox. Even turbulent micro-particle systems, the "deep pools [that] seethe with tumbling waterfalls," speak to harmonies moved rather than transgressed by the power of God. The whole of it thrums with the one uncreated *ratio* of divine Wisdom himself (*Gen. ad litt.* 5.20.41)[49] because the *aeternas rationes*, in their simple unity within Wisdom, have an intrinsic order (*ordo*) that is "hidden from us rather than…lacking to universal nature" (5.21.42).[50] We cannot skip past history to access this order because "our knowledge…depends upon the governance in time of creatures already made, inasmuch as God, in the unfolding of his creatures…is working still" (5.4.10).[51]

Thus, the world is not a collection of static essences defined by distinct forms existing on a different plane. Because there is no distinct world of forms, but only this world or the very mind of God himself, we know *rationes* as distinct only through historical and material existence (*Gen. ad litt.* 5.16.34).[52] A creature's *ratio* is not a functional form obscured by material conditions, but a trajectory that works itself out *through* material conditions and in relation to other *rationes*. When we take up one *ratio*, we take up a knot in the whole tapestry. In this kaleidoscopic theophany of things, their ways, and their histories, "the

---

[48] See also Cavadini, "Augustine and Science," 64. Also, *Gen. ad litt.* 5.21.42; *Literal Meaning of Genesis*, 1:172: "Creatures shaped and born in time should teach us how we ought to regard them. For it is not without reason that Scripture says of Wisdom, that *she graciously appears* to her lovers *in their paths and meets them with unfailing providence* (Wis. 6:17)."

[49] Augustine, *Literal Meaning of Genesis*, 1:171–72: "God moves his whole creation by a hidden power, and all creatures are subject to this movement: the angels carry out his commands, the stars move in their courses, the winds blow now this way, now that, deep pools seethe with tumbling waterfalls and mists forming above them, meadows come to life as their seeds put forth the grass, animals are born and live their lives according to their proper instincts, the evil are permitted to try the just. It is thus that God unfolds the generations that he laid up in creation when first he founded it; and they would not be sent forth to run their course if he who made creatures ceased to exercise his provident rule over them."

[50] Augustine, *Literal Meaning of Genesis*, 1:172.

[51] Augustine, *Literal Meaning of Genesis*, 1:153.

[52] Augustine, *Literal Meaning of Genesis*, 1:167: "*In him we live and move and have our being* [Acts 17:28]; but most creatures…being corporeal, are of a different nature, and our mind is unable to see them in God, [that is,] in the archetypes according to which they were made. [Otherwise,] we should know their number, size, and nature, even without seeing them by means of the senses of our body."

world itself in all its ordered change and movement and in all the beauty it presents to our sight," "bears a kind of silent testimony to the fact of its creation, and proclaims that its maker could have been none other than God, the ineffably and invisibly great, the ineffably and invisibly beautiful" (*Ciu.* 11.4.2).[53]

The contingency of a particular *ratio*—that God did not have to make the creature that way, or express himself under that particular economy—is recapitulated in the contingent conditions under which we encounter it materially. The *rationes* imply and are disclosed in the messiness of nature. The *ratio* of a worm is not captured by some definition such as "fleshy flexible moving linear metabolizer," if that definition does not live in the worm's "silent testimony" by its writhing in *this* physical environment, in *this* soil. The "sub-conceptual" detail of the clumping earth after a light rain and the way it shapes the worm's progress—a behavior adapted *to* that sort of soil—is not inconsequential but is entailed in the *ratio* of that creature.

Here, then, is a fitting metaphysical account of the world's diversity and of the inexpressible interior relations held in our concepts—concepts properly used when creatures and their *rationes* are known in their coherence with one another. Herein is Wisdom apparent; herein, the measure and harmony of the whole draws forth our awe and wonder and praise—all of which, John Cavadini writes, would be denied if that whole were formalistically "reduc[ed] to our rationality" in the sense critiqued by Gilkey and Smith.[54] For an Augustinian metaphysics of creation, the neural network at its best is attuned not to happy accidents but to the *rationes*; the network's mathematics do not refute so much as deepen our notion of the concept as an engagement with those *rationes*.[55]

The metaphysical and epistemological challenge of the neural network has led us not to abandon concepts, but to see them as rooted in the historical outworking of reality through the activities of particular things, with histories and their kinds understood as refractions of a simple and unitary divine Wisdom. What, then, has the network captured when it is trained to distinguish images of plants according to a scientific taxonomy? Philosopher Cameron Buckner writes: "The exact boundaries of each category's manifolds," that is, each category's geography in the data-space,

---

[53]Augustine of Hippo, *City of God,* 432.

[54] Cavadini, "Augustine and Science," 64–65, discussing Augustine, *Gen. ad litt.* 5.22.43.

[55] I do not here have space to deal with networks that discover false correlations or biased shortcuts in data, but will say that such problems do not defeat the claim that, when attuned to data causally linked to nature, the network is attuned to the *rationes* in some measure.

> are inaccessible to networks during training; the "goal" of training a neural network for classification can then be understood as discovering a global output function—composed of individual nodes' activation functions and associated link weights—that can draw boundaries between the manifolds of categories that need to be discriminated.[56]

Certainly, the categories may be determined by the human designer, but a network that maintains robust performance when tested has, in a certain way, accomplished a mapping from the *rationes* (manifested in the world-data presented to the network) to the interests of the network's trainer. It has to *work*; and so it must preserve the *rationes* as much as our own categories do in experience even if not in explicit definition—but this means that it can also distort these *rationes*.

KNOWLEDGE, WISDOM, AND ARTIFACTUAL MEMORY

Having inquired into the neural network's relation to nature's *rationes*, we may now consider its relation to human understanding. The network's outputs are human-designated classifications, categories, and purposes to which the network is trained to map its input data. This is how semantics are attributed to the AI program. The strengths and potential deficiencies of this mapping can make it the subject of a deeper moral-theological reflection:

First, for Augustine, concepts in the human mind are begotten by intentional and moral judgments that, in turn, form the very fabric of our understanding as an engagement with the sub-conceptual web—much like the neural network.

Second, while knowledge (*scientia*) directs these understandings toward our own purposes, true wisdom (*sapientia*) receives the *rationes* contemplatively, allowing them to exceed the scope of any purpose. Always developed for a particular task, AI points beyond itself but, in pointing toward *us* before it points toward the world, it seems unable to transcend a utilitarian frame.

Third, I conclude that, as AI cannot escape the morally infused nature of all human thought, we must develop a "spirituality" of AI wherein we do not permit it to stand between us and the world—lest we remain self-imprisoned in the knowledge of our own designs.

MEANING AS MORAL: VERBUM AND MEMORIA

Every act of understanding involves an act of the will. For Augustine, our acts interpreting natural things, conventional signs, and artifacts all follow the same fundamental sequence: we apprehend something through the senses; we judge it as good (i.e., as real)[57] with

---

[56] Buckner, "Deep Learning," 10. Dashes added for clarity.

[57] Moral evils like murder are "good" only in, say, involving voluntary motion. The act itself forestalls any goodness beyond the bare fact of this motion, in intentionally

respect to something else; then, as we cling to that goodness with our approbation or love, we conceive a "mental word" (*verbum mentis*)—i.e., a conceptual understanding (*Trin.* 9.6.11–9.11.16, 15.10.17–15.11.21).[58] The *verbum mentis* is not a spoken word, nor an autonomous form in Gilkey's sense, but a particular embrace by the mind of some facet of reality according to its *ratio*—an embrace, however, that is shaped by the knower's own assessment of its goodness.[59] For Augustine, in the words of Luigi Gioia, "Intellectual knowledge is not the result of an 'infusion' in our mind of a pre-existing reality, but the production of a new reality."[60]

This *verbum* is truer as it approaches an embrace of the *ratio* as that *ratio* is; and this means that one's own desire and love must conform to reality rather than plucking out from reality only that which is congenial to the stance that one has brought with oneself. Even our recognitions of a narwhal or a "no parking" sign are not neutral because our judgments of meaning issue within the general frame of our cultural, societal, and personal values and position within the world. Every act of understanding entails a moral judgment; habitual moral judgments of this sort form our habit of seeing the world.

Augustine calls the ground of this habitual vision our "memory" (*memoria*).[61] While corporeal things cannot be kept uninterruptedly before the physical eyes, *memoria* makes present the object of the mind's striving, such that God and corporeal objects alike can be present uninterruptedly. The *memoria* is not, however, a movie-screen or data repository (*Conf.* 10.17.26). Rather, it is an implicit knowledge of objects and experiences, a fabric of varyingly accurate *rationes* built up from apprehensions in *verba mentis*. Contained implicitly in this fabric, objects can be said to be present to the mind even without conscious cognition. Desire or love—the will's implicit judgment concerning the thing known—draws the object anew into explicit thought as a *verbum mentis* in the intellect. As Augustine writes in the

---

extinguishing the goodness of one personal life by the agent's ugly inter-personal attempt at absolute domination.

[58] Luigi Gioia writes: "The process of knowledge is set off by desire for the object to be known and is completed only through union with the object known through love" (*The Theological Epistemology of Augustine's* De Trinitate, 200).

[59] John C. Cavadini, "The Quest for Truth in Augustine's *De Trinitate*," *Theological Studies* 58, no. 3 (September 1, 1997): 429–40, doi.org/10.1177/004056399705800302.

[60] Gioia, *Theological Epistemology of Augustine's* De Trinitate, 200.

[61] The texts of Augustine dealing most prominently with *memoria* include: *Conf.* 10; *Trin.* 9, 14, 15.19–20. On the *verbum mentis* see *Trin.* 9.11–12; 15.11.20. See also Nello Cipriani, "Memory," trans. Matthew O'Connell, in *Augustine Through the Ages: An Encyclopedia*, ed. Allan Fitzgerald and John C. Cavadini (Grand Rapids, MI: Eerdmans, 1999); Matthew L. Lamb, "St. Augustine on *Memoria* and *Commemoratio*," in *Philosophy and Theology in the Long Middle Ages*, ed. Kent Emery (Boston, MA: Brill, 2011), 237–47.

Confessions, "I hid in my memory not the images but the realities"—that is, the *rationes* as construed in the *verba mentis* (*Conf.* 10.10.17).

More than permitting implicit presence, the *memoria* is a kind of ground *for* thought, formed *by* thought. One's past apprehensions become the fibers from which one's present intuitive leaps are made and concepts past and present (re)woven (*Trin.* 12.14.23). To use a mathematical analogy, the *memoria* is a set of basis vectors, more or less approaching the true principal components of the vector space that is reality. As built up from the *verba mentis* shaped by the will, the *memoria* constitutes the deep substructure of understanding wherein the *verba* of past and future subsist. Like the palate, the *memoria* is cultivated by the things one tastes attentively and potentiates what and how one is able to taste: past judgments shape the *memoria* and the *memoria* is also the substrate wherein the resulting *verba* are sustained. It is our sensitivity to reality, the primary colors of thought, our way of seeing the world, a habit of mind shaping the judgments that will come readily to us, and a sort of sedimentary aggregate of the *verba* begotten over one's lifetime. If a particular *ratio* is a knot in the tapestry of reality, then *memoria* is a corresponding tapestry of mind from which the *verbum mentis* comes forth. Finally, inasmuch as the will and the affections are susceptible of reformation, the *memoria* is malleable as well.

ARTIFACTUAL MEMORIA: KNOWLEDGE (SCIENTIA) BUT NOT WISDOM (SAPIENTIA)

As an artifact, the programmed computer receives its semantics from the meaning-making intentional frame constituted by the judgments of those who share that frame. Now, if the trained neural network maps the *rationes* of some dataset to categories of interest to the system designer; and if this mapping preserves those *rationes* to the extent that they can be transduced without loss into the designer's moral and conceptual engagement with reality; then the neural network is an artifactual *memoria*, its learned weights preserving the mapping of *rationes*. As with a network sensitive to ridged stalks, this *memoria* is not transparently interpretable in terms of scientific classifications or formal conceptual relationships, but nonetheless it encodes the *rationes* of its input data as shaped by the wills of its designers and users, mapping reality to human interest and utility. Thus, its meaning as used in the world involves both the *verba mentis* that shape the system's architecture and especially its trained outputs, and the moral stances implicit in the goals and purposes to which its users put the AI.[62] When it is read as a standard, taken as a prompt for action, or

[62] This remains the case even for apparently purely "scientific" uses. Weather prediction has goals and valences embedded in it—what we deem important, what is the difference between light and heavy rain, what effects are worth singling out for

contemplated for what it reveals, it thus has ultimate reference not to God (as with the *rationes* of the natural world) but to ourselves.

The artifactual *memoria* therefore subserves what Augustine calls "knowledge" (*scientia*) as distinct from "wisdom" (*sapientia*). *Scientia* apprehends things according to their *rationes* for the sake of "action" in the "good use of temporal things." The moral stances and judgments that beget a right *scientia* build up the "virtues that make for right living" along the way to eternal life. The neural network cannot mirror, however, the higher form of knowledge that is "wisdom" (*sapientia*). *Sapientia* engages the *rationes* not according to their usefulness but as they echo the *aeternas rationes* that are one in divine Wisdom; thus, it reaches toward the "contemplation of eternal things" in God himself (*Trin.* 12.14.21–22).

Whereas the *verba* of *scientia* are begotten by a morally oriented will, *sapientia* enters a transcendent frame because it is begotten by the higher love of "charity" (*caritas*), "poured forth in the heart by the Holy Spirit who is given to us" (Rom 5:5). By charity, one participates in God's own life,[63] and so by a long apprenticeship, the Christian's loves may be brought into this frame so that *scientia* will flow seamlessly into *sapientia* as one refers the goodness of all things to the originating goodness of God, loving them *in* God, with him rather than our temporal purposes being the horizon of their meaning (*Doctr. Chr.* 1.3.3–1.4.4). To be truly wise, in Augustine's sense, is to live from within the life of God according to the self-donative love that is the life of the Trinity, and to know according to Wisdom by finding in each created thing a glimmer of the *aeternas rationes*, which are one in God's eternal Wisdom—i.e., Christ, the second person of the Trinity. *Sapientia*, then, is not simply a matter of having a connected view of things, nor only of knowing the causes of things; it is a configuration of the mind according to God, actualized in a relationship *with* God. In this, one lives fully as God's image by remembering, understanding, and loving the Trinity in direct relationship (*Trin.* 14). By this active participation in God's own life, the "mapping" of human *memoria* becomes a living sign and image not first of the world nor of one's worldly purposes, but of Christ, who is God's own self-knowledge. The wisdom begotten in this *memoria* is a vision beyond words, a contemplation beyond representation, received in

identification; all of these have to do with the human scale of life in the world and the interest that we have in it. We must delineate the concepts else how can it enter our web of meaning? Language translation is a particularly knotty case that I hope to address in a future paper.

[63] David Vincent Meconi, "Augustine's Doctrine of Deification," in *Cambridge Companion to Augustine*, 208–28, universitypublishingonline.org/ref/id/companions/CCO9781139178044A023; Ron Haflidson, "We Shall Be That Seventh Day," in *Deification in the Latin Patristic Tradition*, ed. Jared Ortiz (Washington, DC: Catholic University of America Press, 2019), 169–89.

relationship. This is what it is most fully to eat of the Tree of Life (Prov 3:18; John 17:3; *Gen. ad litt.* 8.5.9–11).

Here we come to what I tentatively propose as a fundamental limitation of the neural network. While *scientia* can be uplifted into *sapientia* within the human mind, I would argue—tentatively—that the interior structure of the neural network, considered as artifactual *memoria*, cannot. This is because, simply, the outputs of the network—when they dictate human action—do so in terms of exterior acts (sell stock), or world-associated categories (thunderstorm). This is why the network's performance is objectively measurable, generating the learning signal by which its weights are adjusted. Such an artifact cannot represent or point to *caritas* because *caritas* is a reality measured not firstly by world-definable ends and actions but by an interior embrace of God as one's friend, even spouse. *Caritas* dictates concrete dispositions in the world but it is not captured by classifications and action decisions. Such a transcendent frame can be declared (i.e., we could train a network to infer and comment upon one's ordering of loves) but it cannot be captured except in terms of its effects in the world. Networks do not have real and subjectively alive relationships in Augustine's sense.

Without a sensitivity to *caritas*, the network cannot become an artifactual *sapientia*. The human observer might reframe the network's meaning beyond its original instrumentality. One might even develop a network to facilitate contemplation of the natural world as a refraction of divine Wisdom. However, as *memoria*—that is, within the intentional frame by which it is trained to map from the world to world-measurable human purposes—it could not be said to represent the *rationes* of created things *as* referring to God. On the part of the human being, the meaning of the network could perhaps be held open to something more, but here it would not be *memoria* but only a sign incomplete in itself because it is unable to accommodate a transcendent frame in the trace of its interior.[64]

THE TWO TREES: OUR CHOICE IN USING AI

It is fitting to conclude this paper by recalling Peter Norvig's suggestion, that we might rightly be satisfied with statistical AI, which "describes what *does* happen" but "doesn't answer the question of *why*," even to the point of bearing no relation to the natural generative processes that give rise to the predicted phenomenon.[65] I suggest that

[64] The question of the network as a predictor and hence a representer of human behavior is intriguing. The love of human beings that frames the meaning of their behavior and their artifacts is ambiguous, almost outwardly incoherent, in that it is shaped both by *caritas* and covetousness, by humble love and the self-defeating autonomy of pride.

[65] Norvig, "On Chomsky." Emphasis original.

such satisfaction would be dangerous, because the opacity of the network's "reasoning" lends itself to becoming a replacement for rather than an invitation to the world, reducing our engagement with the world to the scope of our desires and intentions—to the point that we risk rewriting the world itself as a resource for the accomplishment of our designs, with ourselves rather than divine Wisdom as its ultimate *ratio*. This paper cannot fully expound these familiar themes; here I but gesture to a landscape that demands re-exploration in light of AI.

We deal with the network in terms of outputs for which our own goals are the necessary framing. The network's interior—even as an echo of *memoria*—is recondite. It is manifested first to us by the network's responses to various inputs, somewhat as an animal's instincts are manifested in its behavior. Like these instincts, which must be studied and tested and even then not fully understood, the network—seen from without—suggests its implicit "concepts" but hides them from our view. An animal does not judge the world; it does not theorize about but works *within* the reality with which it interacts. Similarly, the neural network, for all its sophisticated ability to predict data correlations that we might never have imagined, remains in this sense at the level of the animal.

We, on the other hand, ask about reality because it is by judgment that we come to understand. We can see the animal as an invitation to judgment: acting according to its own *ratio*, the animal *is* itself a mapping of the world that we might judge. But the neural network invites our judgment especially because it ultimately concerns us, who have determined the outputs of interest. The network is not a theory or explanation of the world, but *itself* something to be theorized and conceptualized. At best, it may help us to interrogate our own purposes, or it may redirect us (as with the ridged stalks) to the meaning of the world. At worst, it may hide the world from us by hiding its own workings except for its efficacious aid to our own goals.

The interior behavior of artificial neural networks suggests a rich ontology well accounted-for by Augustine's *rationes*. On the other hand, if the opaque network becomes a buffer *between* us and the world, then we risk a very different assumed ontology. The ancient Babylonians, un-wondering masters of data-fitting, sought exacting astrological forecasts but showed no interest in astronomical mechanisms.[66] Their cultural orientation matched their cosmogony, in which the world was fashioned from the bisected carcass of a primordial chaos dragon, slain by her own descendants and held together by ever-

[66] Philip Ball, "Stop Calling the Babylonians Scientists," *The Atlantic*, February 10, 2016, www.theatlantic.com/science/archive/2016/02/babylonians-scientists/462150/; Pearl, "Limitations."

vigilant heavenly guardians.[67] It is not a world that invites science because it is not a world that requites wonder. Redolent of menace, it is not to be understood but only to be controlled.[68]

Our stance of will toward the world shapes our own implicit practical ontology, a reading of the world's *rationes* and a particular actualization of its capacity for meaning. If statistical AI is used as an unexamined instrument of reduction, harnessing the world without understanding, then we shall become practical Babylonians, the instrumental framing of the network dominating our framing of the world itself. As John Cavadini puts it, for Augustine "the sign systems we create are no better than the love in which they were ultimately begotten."[69] A love that values the world merely for its amenability to AI-driven mastery is a love closed to *sapientia*. Such a love will fast decline from *scientia* into mere *superbia* ("pride"), the fatuous science of false autonomy that reduces all to the scope of our perceived desires, so as to live the lie of self-complete dependence on nothing—as if we were gods (Gen 3:5). The artificial neural network can serve our sapiential tasting of the tree of life (although it cannot capture that Wisdom); but, if permitted to delimit our relationship to the world, it will become the Tree of Knowledge, denying to us all that cannot be represented by the instrumental structures by which we have cultivated the network's activity and rendered it intelligible. Is this not the basic dynamic of AI bias? A network tells us what we already "know" because we train it to reduce the world as we do; or it reduces the world in ways we do not notice because our purposes are shaped by the reductive character of our own biases.

What, then, must we do? We must inquire of the world—and we must let the network lead us back to it by inquiring into the network, by striving to understand its working and refusing the easy claim that it bears no relation to the generative processes of nature itself. In that it must reckon in some sense with the *rationes* that have divine Wisdom as their source, a network that cannot accommodate the fullness of the world can still perhaps lead us to it by routes unexpected.

This leads us to the use of artificial intelligence as a spiritual activity. Our behavior and goals are the inescapable framing of the network itself; and so our deployment of these artifacts must imitate God's providential governance of the universe—arranging and further elucidating the *rationes* to yield meanings that they cannot possess simply on their own. As we undertake this godlike activity, will we seek

---

[67] James B. Pritchard, ed., *The Creation Epic* [Enuma Elish], in *The Ancient Near East: An Anthology of Texts and Pictures* (Princeton, NJ: Princeton University Press, 1958), 31–39.

[68] On the political theology of this situation, see Joseph Ratzinger, *"In the Beginning…": A Catholic Understanding of the Story of Creation and the Fall*, trans. Boniface Ramsey (Grand Rapids, MI: Eerdmans, 1995).

[69] Cavadini, "Quest for Truth in Augustine's *De Trinitate*," 436.

greater understanding or only greater efficacy? Let us not be harsh imperators of an unstable world. Instead, let us seek an unveiling of the dynamics of creation for use according to their intrinsic goodness and meaning. The network maps natural things to conventional meanings, but if we return to the world, we prevent those meanings from merely signifying ourselves. The right use of AI does not depend merely on the architecture of our systems, nor even on the ethics that we attempt to embed in them, but on the ultimate stance of will that we adopt—be it *superbia* or *caritas*, unto a false knowledge or a true *scientia* and, finally, wisdom. This is the challenge of AI, our moral framing of which will determine what of reality we permit ourselves to see. M

Jordan Joseph Wales is Associate Professor and the John and Helen Kuczmarski Chair in Theology at Hillsdale College. His scholarship—which has appeared in the journals *AI & Society*, *Augustinian Studies*, and *Cistercian Studies Quarterly*, among others—focuses on early Christianity as well as contemporary questions relating to theology and artificial intelligence. He currently is working on two books—one on the theology of Gregory the Great and another on AI and theology. Holding degrees in Engineering (BS), Cognitive Science (MSc), and Theology (MTS, PhD), he is an advisor to the Holy See's new Center for Digital Culture, under the Pontifical Council for Culture; and he is an affiliated scholar with the Centre for Humanity and the Common Good at Regent College, University of British Columbia.

Journal of Moral Theology, Vol. 11, Special Issue 1 (2022): 182–211

Theological Foundations for Moral Artificial Intelligence[1]

Mark Graves

KEY IDEAS FROM MORAL THEOLOGY CAN help make AI compatible with human morality by guiding the integration of disparate approaches to AI development toward a morally good end. As AI becomes more pervasive in society, humanity would benefit from AI development incorporating a theological anthropology that can guide AI's interdisciplinary construction and characterize its historically contextualized moral norms. As an initial foray into development of an integrative framework, I describe an AI system that could plausibly be constructed with effort comparable to other major AI initiatives, and that would have the capacity to consider itself as a moral actor (a precursor to moral agency).[2] Constructing such a system would open up new possibilities for moral AI, enable sophisticated modeling of human morality, and lead to new insights into ethics and moral behavior. Closer at hand, my proposal identifies issues in AI and morality that require both computational and ethical expertise to resolve and are not well known and understood across the necessary disciplines.

As I use the term, "moral AI" can navigate the moral dimension of its world and predict the moral consequences of its actions. To do so

[1] The initiation of the project described by this manuscript was made possible through a fellowship funded by John Templeton Foundation through St. Andrews University and the University of Notre Dame Center for Theology, Science & Human Flourishing with Celia Deane-Drummond. My project benefited from interactions through St. Andrews and at Notre Dame, especially conversations with Darcia Narváez, Emanuele Ratti, Tim Reilly, and Adam Willows and specific topics of the paper were informed by early conversations with Jean Porter, Bill Mattison, and Walter Scheirer. Thanks to Bob Lasalle-Klein, Rene Sanchez, José Sols Lucia, Pat Lippert, and other members of the John Courtney Murray group for comments and suggestions on an earlier draft. Andrew Porter was very helpful in identifying an early direction. A prior version of this article benefited from discussion at a Pacific Coast Theological Society meeting, especially comments by Brian Green, Katy Dickinson, Bob Russell, Ted Peters, Koo Yun, Braden Molhoek, Kenn Christianson, and John LaMuth. The article also significantly benefited from comments by two anonymous reviewers and the special issue editors.

[2] As explained later in this article, the difference between "actor" and motivated "agent" draws upon Dan P. McAdams, "The Psychological Self as Actor, Agent, and Author," *Perspectives on Psychological Science* 8, no. 3 (2013): 272–95.

it must conceptualize its natural, social, and moral world and reckon itself within those worlds.[3] When an AI reckons itself: (1) as a causal actor, it can engage the natural world; (2) as a sociotechnical actor, it can develop communicative relationships with others in its social world; and (3) as a moral actor, it can evaluate the ethical consequences of its actions in its moral world. An interdisciplinary construction of moral AI depends upon insights into morality and AI development, and can contribute to both as well as beneficial incorporation of AI technology into society. Many of the above words such as "moral," "conceptualize," "actor," "reckon," etc., we typically reserve for the behaviors of self-conscious agents like humans, and while I do not rely on that interpretation here, I leave open the possibility that AI might someday attain that status.[4] Several of these terms will be more fully elucidated later on, with attention to their formulation separate from assumptions of consciousness.

A number of disciplinary perspectives contribute to the development of moral AI. Computer scientists often recognize the need for ethical AI, and incorporating ethical principles into AI development, such as fairness, is an active AI research area.[5] Social scientists have studied human interaction with AI including people's tendency to anthropomorphize AI and differences in trusting AI versus humans.[6] Collaborations between philosophers, ethicists, and others have

[3] For evidence of neural networks exhibiting concept-like functioning, see Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah, "Multimodal Neurons in Artificial Neural Networks," *Distill*, 2021, distill.pub/2021/multimodal-neurons/.

[4] For differing opinions on whether AI can have self-consciousness or interiority, see Brian P. Green, Matthew J. Gaudet, Levi Checketts, Brian Cutter, Noreen Herzfeld, Cory Lebrecque, Anselm Ramelow, OP, Paul Scherz, Marga Vega, Andrea Vicini, and Jordan Joseph Wales, "Artificial Intelligence and Moral Theology: A Conversation," *Journal of Moral Theology* 11, Special Issue 1 (Spring 2022): 13–40.

[5] Stuart Russell, Daniel Dewey, and Max Tegmark, "Research Priorities for Robust and Beneficial Artificial Intelligence," *AI Magazine* 36, no. 4 (December 31, 2015): 105–14, doi.org/10.1609/aimag.v36i4.2577; Pat Langley, "Explainable, Normative, and Justified Agency," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33 (2019): 9775–79, doi.org/10.1609/aaai.v33i01.33019775; Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi, "Fairness and Abstraction in Sociotechnical Systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19 (New York: Association for Computing Machinery, 2019), 59–68, doi.org/10.1145/3287560.3287598; and Donghee Shin and Yong Jin Park, "Role of Fairness, Accountability, and Transparency in Algorithmic Affordance," *Computers in Human Behavior* 98 (2019): 277–84, doi.org/10.1016/j.chb.2019.04.019.

[6] Arleen Salles, Kathinka Evers, and Michele Farisco, "Anthropomorphism in AI," *AJOB Neuroscience* 11, no. 2 (April 2, 2020): 88–95, doi.org/10.1080/21507740.2020.1740350; Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H. de Vreese, "In AI We Trust? Perceptions about Automated Decision-Making by Artificial Intelligence," *AI & Society* 35, no. 3 (2020): 611–23, doi.org/10.1007/s00146-019-00931-w.

identified ethical principles and practices for incorporating AI predictions and other results into social structures.[7] Machine ethicists have clarified the need for explicit characterizations of ethics and the need to reconcile differences between what distinct duties (or other value frameworks) might require.[8] Theologians have begun examining AI in the context of theological anthropology, and elsewhere in this volume, moral theology.[9] Collaborative engagement on the development of moral AI can prescribe key components for AI development and guide ongoing efforts to incorporate ethics into AI.

Moral theologians can help construct a framework to integrate technical, social, and ethical contributions on AI with scientific, scholarly, and normative insights into human society. Although differences among ethical theories, schools of thought, and religious traditions are legion, I agree with ethicist Susan Anderson that enough consensus on ethical thought exists to guide construction of moral AI.[10] However, constructing moral AI is a normative process, not a descriptive one, and although what exists in human morality is an important aspect of developing moral AI, building an AI system with moral judgment and behavior requires reasoning about moral normativity in a moral actor with radically different embodiment and socialization. AI developers

---

[7] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena, "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and Machines* 28, no. 4 (2018): 689–707, doi.org/10.1007/s11023-018-9482-5; Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal, "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," *Science and Engineering Ethics* 26, no. 4 (August 1, 2020): 2141–68, doi.org/10.1007/s11948-019-00165-5.

[8] Michael Anderson and Susan Leigh Anderson, *Machine Ethics* (Cambridge: Cambridge University Press, 2011); Wendell Wallach and Peter Asaro, *Machine Ethics and Robot Ethics* (New York: Routledge, 2017); Susan Leigh Anderson, "Machine Metaethics," in *Machine Ethics*, ed. Michael Anderson and Susan Leigh Anderson (Cambridge: Cambridge University Press, 2011), 21–27.

[9] Noreen L Herzfeld, *In Our Image: Artificial Intelligence and the Human Spirit* (Minneapolis, MN: Fortress, 2002); Anne Foerst, *God in the Machine: What Robots Teach Us about Humanity and God* (New York: Dutton, 2004); William F. Clocksin, "Artificial Intelligence and the Future," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 361, no. 1809 (2003): 1721–48, doi.org/10.1098/rsta.2003.1232; Russell C. Bjork, "Artificial Intelligence and the Soul," *Perspectives on Science and Christian Faith* 60, no. 2 (2008): 95–102; Andrew Peabody Porter, "A Theologian Looks at AI," in *2014 AAAI Fall Symposium Series*, 2014.

[10] Anderson, "Machine Metaethics." Practical issues that would require theoretical ethical nuance also require significant immersion in technology development. Philosopher of technology ethics Shannon Vallor makes a similar point on consensus. See her *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (New York: Oxford University Press, 2016), doi.org/10.1093/acprof:oso/9780190498511.003.0001.

often have moral intuitions grounded in a rich intellectual tradition but lack the historical and philosophical knowledge and expertise to make those intuitions explicit for machine ethics; and ethicists typically lack sufficient insight into rapidly developing technologies to identify detailed social and moral implications before technical development has progressed past the point of immediate relevancy. Moral theologians can help bridge that gap with an integrative framework for moral AI within which other disciplines can dialogue and collaborate.

The Interdisciplinary Challenge: Snow's "Two Cultures" Problem

A challenge to interdisciplinary investigation of moral AI is the relatively non-overlapping educational training of computer scientists (and engineers) and moral theologians (and philosophers and ethicists), which severely limits the construction of robust theories incorporating both advanced technical understanding and scholarly insight. One can trace recognition of the challenge to C. P. Snow's identification of two cultures separating science and the humanities.[11] Differences in the presumed background knowledge and trained methodologies hinder dialogue between scientists and scholars, and sophisticated theories in one discipline may include assumptions considered naive by the other. Ian Barbour and others have previously studied challenges to dialogue between theology and natural science, and studying AI morality can draw upon those lessons. Advances also require integrating that academic discourse with its related technology and ethics dialogue, previously viewed primarily as applications of science and theology, respectively.[12] In the case of AI morality, this integration reverses the previously noted distinction between theoretician and practitioner. For the specific technological application of interest is an engineered system that threatens to replicate the experience and intellectual expertise previously presumed the exclusive purview of scientists and theologians.[13] One must also incorporate the social sciences

[11] C. P. Snow, *The Two Cultures and the Scientific Revolution* (New York: Cambridge University Press, 1959).

[12] Ian G. Barbour, *Religion and Science: Historical and Contemporary Issues* (San Francisco: Harper, 1997); and Ian G. Barbour, *Ethics in an Age of Technology* (San Francisco: Harper, 1993).

[13] Joe Dysart, "The Writing Is on the Wall for Artificial Intelligence," *Research-Technology Management* 62, no. 6 (2019): 8; Beta Writer, *Lithium-Ion Batteries: A Machine-Generated Summary of Current Research* (Springer International, 2019), www.springer.com/us/book/9783030167998; Mark Graves, "AI Reading Theology: Promises and Perils," in *AI and IA: Utopia or Extinction?*, *Agathon* 5 (2018); and Xin He, Kaiyong Zhao, and Xiaowen Chu, "AutoML: A Survey of the State-of-the-Art," *Knowledge-Based Systems* 212 (January 5, 2021): 106622, doi.org/10.1016/j.knosys.2020.106622. Because AI fundamentally relates to human experience and mental processing in a way no previous technology has, it depends in a novel way upon and can impact every field that studies or relies upon human cognition. Studying AI morality not only requires innovative integration of humanities with

as they identify social structures that AI impacts and disrupts as well as explain the human psychology that AI purports to replicate partially and with which AI must often interact. The social sciences are also needed because philosophers and computer scientists like John Searle, Hubert Dreyfus, and Brian Cantwell Smith convincingly identify certain knowledge, phenomenological engagement, and commitments to the world as missing in AI but do not appear to fully appreciate the relevant and nuanced contributions to those mental capacities by sociology of knowledge and social and developmental psychology, even for humans.[14] The interdisciplinary challenge is addressed through a collaborative framework for moral AI development that can integrate the discipline-specific theories and shift efforts from loose discussion and dialogue to something that focuses and constrains contributors sufficiently to impact theories and practices from other contributing disciplines.

Moral AI raises many questions of personhood not addressable in a single article, and some assumptions must be made with respect to AI's cognitive capabilities, moral agency, phenomenological consciousness, and moral continuity with humans.[15] Possible AI cognitive capabilities can variously refer to the equivalent of: (1) an artifact such as a calculator or computer, (2) an intelligent non-human animal, (3) that new intelligent animal-like "species" plus language and culture, or (4) also include a degree of self-awareness and reflection, most similar to modern humans.[16] Other options are possible as well. Here I aim to clarify how an AI beginning with intelligence of a non-human animal can add the capability to participate in the human social world, which enables better characterization of the necessary preconditions for self-reckoning as a foundation for self-awareness and reflection.[17]

---

natural and social sciences, it can also require examining the presumptions and historical accidents that led to their separation.

[14] John R. Searle, *Minds, Brains, and Science* (Cambridge, MA: Harvard University Press, 1984); Hubert L. Dreyfus, "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian," *Philosophical Psychology* 20, no. 2 (2007): 247–68; and Brian Cantwell Smith, *The Promise of Artificial Intelligence: Reckoning and Judgment* (Cambridge, MA: MIT Press, 2019). See also Porter, "A Theologian Looks at AI."

[15] The use of "AI" as an entity, instead of a research field, presumes a not-yet-existent level of cohesion and generalizability among the outputs of that field, which requires additional integrative work, such as proposed here.

[16] Comparing cognition between humans and AI is possible because the fields of AI and cognitive psychology have informed each other's development within the broad umbrella of cognitive science, resulting in compatible scientific characterizations between human and AI cognition, though their mechanisms, embodied realization, and phenomenological concerns differ substantially. See George A Miller, "The Cognitive Revolution: A Historical Perspective," *Trends in Cognitive Science* 7, no. 3 (2003): 141–44.

[17] In this usage, self-reckoning is a foundation for self-awareness, but the self lacks awareness of itself as a "knower."

Moral agency often implies a high degree of autonomy, though AI could have restricted (e.g., safe) agency; exist in a way so its "free will" is "compatible" with an otherwise deterministic foundation; or result from humans giving it equivalence to agency in a sociotechnical system, such as of a judge, loan officer, or corporate executive, even though the AI technology lacks intrinsic agency.[18] Common to all these types of moral agency is the capacity of AI for moral attention and interpretation and ultimately the ability to judge the impacts of its own decision making. I focus on AI interpreting its world in a way that admits moral decisions and action and includes recognition of its own actions, without requiring those decisions and actions to be motivated or autonomous. Considering the range of AI's relationships to its "self" from none through self-reckoning to full phenomenological consciousness and reflection upon its inner life, I target self-reckoning as AI perceiving its own existence in its world, but not necessarily any greater awareness of itself or its interior processing. I argue that an AI with these cognitive and self-reckoning capacities engaging a human social world through language and attending to value-laden and normative interpretations suffices as a foundation for considering AI's moral continuity with humans in that world.[19]

A Framework for Moral Theology and AI Research

In this article, I propose an initial framework for drawing moral theologians into the multifaceted, integrative discourse on moral AI. The article unfolds in two main parts. First, a theological foundation for moral AI requires something like a secularized theological anthropology. The "anthropology" characterizes the natural, social, and moral aspects of an AI that exists in a world with humans, sin, and grace and focuses on what is needed to characterize such a social and moral entity (though without directly attributing sin or grace to AI). Critiques of current approaches to AI identify limitations to AI's more anthropological development, and I respond by adapting Donald Gelpi's theological anthropology for moral AI to emphasize the AI's

[18] John McCarthy, "Free Will—Even for Robots," *Journal of Experimental & Theoretical Artificial Intelligence* 12, no. 3 (July 2000): 341–52, doi.org/10.1080/09528130050111473; Riccardo Manzotti, "Machine Free Will: Is Free Will a Necessary Ingredient of Machine Consciousness?," *Advances in Experimental Medicine and Biology* 718 (January 1, 2011): 181–91, doi.org/10.1007/978-1-4614-0164-3_15; Paul N. Edwards, "Infrastructure and Modernity: Force, Time, and Social Organization in the History of Sociotechnical Systems," in *Modernity and Technology*, ed. Thomas J. Misa, Philip Brey, and Andrew Feenberg (Cambridge, MA: MIT Press, 2003), 185–226; and Selbst, Boyd, Friedler, Venkatasubramanian, and Vertesi, "Fairness and Abstraction in Sociotechnical Systems."

[19] Although greater capacities would be needed for moral agency, full moral autonomy, or moral equivalency with humans, I claim these capacities suffice for interdisciplinary dialogue about AI meaningfully considered to be moral, and with a more active role than a moral patient.

moral conceptualization and self-reckoning in a casual, social, and moral world.[20] Gelpi's anthropology has a metaphysics rooted in experience, based upon C. S. Peirce's and Josiah Royce's objective idealism, and this provides theological grounding for AI's interpretive experience. To extend the anthropology for moral AI, I: (1) characterize an AI self as a moral actor that experiences its world; (2) use systems theory to organize an AI's interpretive experience of its natural, social, and moral world; (3) situate AI social apprehension within Ignacio Ellacuria's historical reality (with moral implications); and (4) adapt Thomistic ideogenesis to characterize an AI conceptualization of its (interpreted) reality in terms of moral norms. Moral norms refer here to what is modeled as normative by the AI, such as moral principles, Ross's *prima facie* duties, utilitarian preferences, proxies for human flourishing (or safety), or virtues.[21]

In the second part, insights from the extended anthropology lead to a proposal for developing moral AI. In the proposed system, moral AI's interpretive experience is characterized by five levels of models, which draw upon systems theory to characterize the AI's encounter with an external world, and five corresponding stages of self-reckoning, where the AI models itself. The multi-faceted, multi-level characterization also defines a framework that identifies the broad disciplinary needs that arise from the attempt at moral AI and a need for collaboration between moral theologians, ethicists, philosophers, social scientists, and computer scientists. The implications of the modeling are then briefly examined with respect to practical wisdom (*phronesis*) as an essential capability for moral AI.

AI Theological Anthropology

Some AI researchers recognize the need for AI to engage its natural and social world in order to develop further and fulfill its promise instead of its perils. Brian Cantwell Smith argues AI must distinguish reality from its representation and commit not just to its representations but to that to which its representations point.[22] Acknowledging Hubert Dreyfus's Heideggerian critique that AI is unable to grasp reality because symbol processing and representations cannot connect experience with existence, Cantwell Smith draws attention to the process that leads from a phenomenological encounter with reality to the

[20] Smith, *The Promise of Artificial Intelligence*; Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (New York: Penguin, 2019); Donald L Gelpi, *The Gracing of Human Experience: Rethinking the Relationship between Nature and Grace* (Collegeville, MI: Liturgical Press, 2001).

[21] Anderson, "Machine Metaethics"; Russell, *Human Compatible*; Mark Graves, "Shared Moral and Spiritual Development among Human Persons and Artificially Intelligent Agents," *Theology and Science* 15, no. 3 (2017): 333–51, doi.org/10.1080/14746700.2017.1335066.

[22] Smith, *The Promise of Artificial Intelligence*, chaps. 7, 12.

distinction between objects required for AI representation.[23] Additionally, Stuart Russell extends Nick Bostrom's philosophical argument that superintelligent AI poses an existential risk to humanity by identifying problematic assumptions in AI research and plausible future improvements in AI sufficient for uncontrollable AI advancement.[24] Rather than halt AI development, Russell argues for developing beneficial AI that identifies human preferences and attempts to maximize those utilitarian preferences with altruism and humility, specifically acknowledging the intrinsic uncertainty in accurately identifying human preferences.[25] Although not identified as such, both researchers point toward the construct of experience as key to developing AI that would have more general capabilities than the narrow and fragile applications currently available and could engage its natural and social world in an ethical way.

Three philosophical perspectives on human experience relevant for modeling AI experience are Continental phenomenology, Thomistic anthropology, and the objective idealism of pragmatism. Continental phenomenology (especially Merleau-Ponty and Heidegger) separates the experience of reality from reality to examine the former and thus provides a focus on subjective awareness that Cantwell Smith, Russell, and others have identified as needed for AI. Thomistic philosophy presumes an objective account of nature compatible with its medieval understanding of the world, which reconciles well with experience of a virtual world and the assumptions of objectivity influential on engineering and the natural sciences. However, the philosophical presumption of subjectivity by Continental philosophy does not guide engineers trying to construct something like subjectivity in machines; although the assumption of universal essences underlying Thomistic philosophy corresponds surprisingly well to presumptions of early AI knowledge representation systems, it captures poorly the evolutionary processes of the natural world, the social construction of knowledge, and contextualized morality. The objective idealism of pragmatic philosophy addresses these limitations for AI. With respect to Thomism, C. S. Peirce incorporates evolutionary processes into his logical metaphysics, thus adding evolution to an Aristotelian-influenced metaphysics, and Josiah Royce further extends Peirce's semiotic philosophy into the social, moral, and spiritual realm, which adds social and moral contextualization.[26] In addition, the pragmatist George Herbert

[23] Dreyfus, "Why Heideggerian AI Failed"; Smith, *The Promise of Artificial Intelligence*, chap. 3.
[24] Russell, *Human Compatible*, chaps. 2-3; Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014).
[25] Russell, *Human Compatible*, chaps. 7, 9.
[26] Kelly A. Parker, *The Continuity of Peirce's Thought* (Nashville: Vanderbilt University Press, 1998); Josiah Royce, *The Problem of Christianity. Lectures Delivered*

Mead changes the locus of personhood from subject or soul, as in Continental and Thomistic philosophy respectively, to the "self" as a social process, thus identifying social construction of subjectivity.[27] Although pragmatism serves as the foundational philosophical framework, a pragmatic understanding of interpretive experience is strengthened by Continental and Thomistic contributions on subjectivity and objectivity, specifically with respect to historical (and political) reality and conceptualization of moral norms.

Pragmatic Experience of Reality

Pragmatically, experience consists of encounter and interpretation.[28] As subject, one encounters one's world, and then interprets one's experience into objective categories. Subjectivity occurs at the nexus of encounters and is defined by those natural and social experiences. Interpreted "objects" are not *a priori* universals, but socially constructed with others in society (and through history and language). Without the sensory encounter, an overly rational interpretation reduces objective idealism to subjective idealism and loses the connection to the real world required by scientific study. Setting to one side possible revelatory experiences, these "others" have historically always been human, but now other precursors to persons are entering into society.[29]

Mead identifies the locus of personhood or "self" as a social process created by interactions within a group or society.[30] The individual social self initially appropriates society's shared values and ideals then, as it develops, interiorizes the social environment in which it lives, and finally begins transforming society through its relationships. AI currently appropriates society's shared values (including those with harmful effect) but does not yet interiorize the social environment in which it lives.[31] As the human "self" incorporates and responds to its social relationships, its reflective character makes it both subject and object, and its communication creates self-awareness. Although foundational for social psychology, the identification of the self as subject and object has not been sufficiently incorporated into dialogue

at the Lowell Institute in Boston, and at Manchester College, Oxford (New York: Macmillan, 1913).

[27] George Herbert Mead, *Mind, Self & Society from the Standpoint of a Social Behaviorist* (Chicago: University of Chicago Press, 1934).

[28] Denis Edwards, *Human Experience of God* (New York: Paulist, 1983); John Edwin Smith, *Experience and God* (New York: Oxford University Press, 1968).

[29] Mark Coeckelbergh, "Robot Rights? Towards a Social-Relational Justification of Moral Consideration," *Ethics and Information Technology* 12, no. 3 (2010): 209–21.

[30] Mead, *Mind, Self & Society*.

[31] There are computational social models, but they are not yet compatible with natural language processing (NLP) deep learning models appropriating social values and biases. The early AI researcher Allen Newell does identify the Social band in *Unified Theories of Cognition* (Cambridge, MA: Harvard University Press, 1990).

between AI engineering and the humanities. If AI begins with a self that experiences its natural and social world, the question arises: What would make it moral? Advances in AI cognitive architecture and integration among methods and technologies would be required to construct such a foundation but are currently plausible given current technology and effort. Can moral theology construct the theories needed to guide such AI development in a moral direction before such AI exists?

To relate Mead's social self to the level of "self" targeted here for moral AI, a distinction from personality psychology is helpful. Dan McAdams studies the formation of identity and identifies three levels of its variation and development in personality: dispositional traits, which are fairly stable through adulthood; characteristic adaptations, which include beliefs and desires and vary throughout one's life; and narrative identity, which comprises the stories one constructs to give one's life a sense of unity and purpose. He summarizes these developmentally as self as actor, agent, and author.[32] Simplistically, dispositional traits may depend upon early childhood development and other social and genetic factors forming the core of one's self. Conversely, characteristic adaptations are more circumstantial and subjective, depending upon one's social, historical, and cultural context as it influences how one apprehends and responds to reality. As for narrative identity, adults form stories about themselves that give meaning and coherence to their behavior over time. One's story is affected by one's dispositions, circumstances, and one's goals and aspirations. The realization that the "self" develops over time (in a historical-social context) helps explain the limitations of considering the essential locus of a person as an "atomic" subject or soul.[33] In addition, McAdams's distinction between social actor, motivational agent, and autobiographical author specifies potential stages for AI development. Although how the human self develops remains an open area of psychological research, McAdams's model suffices to demonstrate that one cannot obtain AI self-awareness and narrative identity solely from building

[32] McAdams, "The Psychological Self as Actor, Agent, and Author"; Dan P. McAdams, "Narrative Identity: What Is It? What Does It Do? How Do You Measure It?," *Imagination, Cognition, and Personality* 37, no. 3 (2018): 359–72, doi.org/10.1177/0276236618756704.

[33] The neuroscientific correlates of human self-awareness are the subject of active research, but social scientists since Mead have examined the necessity of society in defining one's self, and moral identity appears a significant factor in human moral action. Sam A. Hardy and Gustavo Carlo, "Moral Identity: What Is It, How Does It Develop, and Is It Linked to Moral Action?," *Child Development Perspectives* 5, no. 3 (2011): 212–18, doi.org/10.1111/j.1750-8606.2011.00189.x; Darcia Narváez and Daniel K. Lapsley, eds., *Personality, Identity, and Character: Explorations in Moral Psychology* (Cambridge, MA: Cambridge University Press, 2009); L. J. Walker, "Moral Personality, Motivation, and Identity," in *Handbook of Moral Development*, ed. Melanie Killen and Judith G. Smetana (London: Routledge, 2014), 497–519.

dispositional traits (like in symbolic AI) or characteristic adaptations (like in statistical machine learning), but that both of these aspects of the self must engage social reality to begin to form the substrate for a self.[34] A first step, undertaken in this article, is for AI both to act in a social context and to reckon itself as an actor in that reality.[35] The proposed AI self as actor would thus initially respond stably in a social context but lack the motivation and desires to change how it apprehends reality. Orienting those actions in a moral direction requires the ability for AI to interpret its natural, social, and moral world.

As a theological foundation for an AI moral self, the Jesuit theologian Donald Gelpi's theological anthropology suffices for relating an AI self to reality. As a metaphysical foundation for his anthropology, Gelpi extends Peirce's phenomenological metaphysics with Alfred North Whitehead's metaphysical process of an emerging self to develop a metaphysics of experience.[36] Gelpi refines his experiential metaphysics by drawing upon Mead's construct of social self, to develop a theological anthropology of the autonomous, social, sentient being that experiences the world and develops through decision-making. For Gelpi, decision-making occurs within an evaluative process that results in taking on habits or tendencies, which then become the foundation for one's future decision-making.[37] In Peirce's semiotic metaphysics, interpretation is fundamental, and Gelpi's theological anthropology considers general interpretive capacity as capable of receiving grace in humans. This nexus of dispositions—the human self—experiences reality by interpreting what it encounters. By providing a metaphysical foundation for an experiential self, Gelpi provides ample grounding for considering the particular case of an AI self.[38] To build upon Gelpi's metaphysical and anthropological foundation, it suffices here to simply require that the AI system have the

[34] This extends Brian Cantwell Smith's critical examination by suggesting AI needs to engage not only the natural world but also social reality (Smith, *The Promise of Artificial Intelligence*).

[35] Depending upon how "self" is defined, this would form something like a proto-self without the narrative identity needed for autobiographical consciousness. In Damasio's theory of consciousness, the proposed system is analogous to his protoself with a foundation for core consciousness but may lack the commitment to self which, for humans, is grounded in emotions (Antonio Damasio, *Self Comes to Mind: Constructing the Conscious Brain* [New York: Random House, 2010]).

[36] Gelpi, *The Gracing of Human Experience*.

[37] Metaphysically, the "evaluation process" builds upon C. S. Peirce's category of Firstness, "decision-making" builds upon his category of Secondness, and habits or "tendencies" build upon his category of Thirdness. See Gelpi, *The Gracing of Human Experience*, 153; Parker, *The Continuity of Peirce's Thought,* 113–16; Charles S. Peirce, *Collected Papers* (Cambridge, MA: Belknap, 1960), vol. 1, § 24–26.

[38] For connection between Gelpi's self and cognitive neuroscience (in the context of neo-Thomistic nature and grace), see Mark Graves, "Gracing Neuroscientific Tendencies of the Embodied Soul," *Philosophy and Theology* 26, no. 1 (2014): 97–129, doi.org/10.5840/philtheol20143125.

ability to learn from its decisions in a way that affects future decision making, which is a general feature of most machine learning systems.[39] Although Peirce and Gelpi emphasize the continuity of those human interpretations with the interpretive dispositions of reality, for interdisciplinary development of moral AI, these interpretive dispositions of experience require further organization. Although Gelpi describes a "self" useful for AI, work is needed to identify *how* to construct an AI self, which I also claim would be a precursor to something like AI subjectivity or phenomenological awareness.

Five Levels of Interpretive Experience

Beginning in the 1940s with the seminal work of Ludwig von Bertalanffy, systems theory has attempted to develop a general theory to organize natural and social phenomena based upon patterns and principles common across a range of disciplines.[40] Although an ultimate systems theory of everything remains elusive, systemic principles have proven effective in a variety of fields from biology through clinical psychology to economics and organizational management as well as computer science. These principles' unifying organization supplies an integrated perspective on natural and social sciences sufficient for the present purpose, even though specialized theories may prove more effective in distinct specific areas.

In general systems theory, von Bertalanffy organizes scientific disciplines and systems into four levels based on physical, biological, psychological/behavioral, and social scientific disciplines to discover general rules about systems that cross those levels.[41] Many others take similar approaches, and Arthur Peacocke organizes his own part-whole hierarchies of nature into four similar levels of focus based upon A. A. Abrahamsen's distinctions between the physical world, living organisms, the behavior of living organisms, and human culture.[42] The contemporary philosopher of science and religion Philip

[39] Gelpi's attentiveness to the dispositional nature of the emerging self allows us to incorporate a teleological element in AI development that, without recourse to universals, still supports the development of virtue, and therefore an AI virtue ethic. See Mark Graves, "Habits, Tendencies, and Habitus: The Embodied Soul's Dispositions of Mind, Body, and Person," in *Habits in Mind: Integrating Theology, Philosophy, and the Cognitive Science of Virtue, Emotion, and Character Formation*, ed. Gregory R. Peterson, James van Slyke, Michael Spezio, and Kevin Reimer (Leiden: Brill, 2017).

[40] Ludwig von Bertalanffy, *General System Theory: Foundations, Development, Applications* (New York: G. Braziller, 1969).

[41] Ludwig von Bertalanffy, *Perspectives on General System Theory: Scientific-Philosophical Studies* (New York: G. Braziller, 1975), 5–8, 30–32.

[42] W. Bechtel and A. A. Abrahamsen, *Connectionism and the Mind* (Oxford: Blackwell, 1991), 256–59; Arthur Robert Peacocke, *Theology for a Scientific Age: Being and Becoming—Natural, Divine, and Human* (Minneapolis: Fortress, 1993), 215; Arthur Robert Peacocke, *God and the New Biology* (London: Dent, 1986); Mark Graves,

Clayton suggests an additional level of spiritual or transcendent activity, which emerges from mental (and cultural) activity and would add a fifth level to the systems model.[43] In alignment with a Thomistic anthropology, von Bertalanffy's biological level corresponds to Thomistic vegetative powers; his psychological/behavioral level maps well to Thomistic sensitive powers; and the separation between social/cultural and transcendent levels distinguishes processes that are combined within the Thomistic rational power. Historical and linguistic activity occurs at the social/cultural level, and the resulting presumed universals define the transcendent level. Rather than treat universals as occurring in a separate realm—e.g., the Mind of God (*nous*)—the analogues for universals occur in the transcendent level, similar to how historically separated dualist realms of *élan vital* or *res cogitans* are now well characterized by systems theory as biological and psychological levels, respectively.[44]

Although von Bertalanffy developed systems theory to organize the scientific study of reality, here it is used to characterize AI experience of reality. This organizes AI interpretations of reality into multiple levels of models.[45] Borrowing from human experience, five levels of interpretation would be models of (a) spatial (or virtual) and temporal extent in physical objects; (b) biological processes; (c) sensation and animation typified by most animals; (d) social relations with expressiveness and meaning of symbolic language as a tool for conceptualization and communication; and (e) moral and spiritual concerns and capacities.[46] These interpretive levels suggest an organization for

---

Mind, Brain, and the Elusive Soul: Human Systems of Cognitive Science and Religion (Burlington, VT: Ashgate, 2008), chap. 2.

[43] Philip Clayton, *Mind and Emergence: From Quantum to Consciousness* (New York: Oxford University Press, 2004); Mark Graves, "The Emergence of Transcendental Norms in Human Systems," *Zygon* 44, no. 3 (2009): 501–32.

[44] Elsewhere, I use Terrence Deacon's emergent dynamics to describe how the transcendent-level processes relate to classical universals, such as transcendentals of Truth, Beauty, and the Good. See his "Emergence: The Hole at the Wheel's Hub," in *The Re-Emergence of Emergence*, ed. Philip Clayton and Paul Davies (Oxford: Oxford University Press, 2006), 111–50; Graves, "The Emergence of Transcendental Norms in Human Systems."

[45] The shift to models draws upon both philosophy of science (as modeling external reality) and cognitive psychology (for mental modeling). See Michael Weisberg, *Simulation and Similarity: Using Models to Understand the World* (New York: Oxford University Press, 2013); Philip Nicholas Johnson-Laird, *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness* (Cambridge, MA: Harvard University Press, 1983); Lorenzo Magnani and Claudia Casadio, eds., *Model-Based Reasoning in Science and Technology* (Cham: Springer, 2016).

[46] In a narrow sense, this organization supports my argument that the capacity to represent moral norms sufficient for addressing conflicts depends upon conceptualization using symbolic language to interpret animal-like phenomenological encounters, and that a proto-self sufficient to reckon oneself as actor in a social realm would enable that moral capacity. My broader claim of theological relevance also depends upon the

moral AI systems and a staged taxonomy of AI systems that could be incrementally built before making an AI that seems like a full person to us. This organization must not only model AI's external reality, it must capture AI's reckoning of itself in that reality which, as discussed later, would correspond to itself as a causal, social, and moral actor.[47] With systems theory organizing an AI's interpretive experience, we turn to expanding the subjective and phenomenological and then the objective and conceptual dimensions of that experience.

Apprehension of Social-Historical Reality

Drawing upon Continental philosophy, Dreyfus used Heidegger's characterization of human existence to identify the disconnect between symbolic approaches to AI and the engagement with reality needed to meet its goals.[48] Cantwell Smith extends and contrasts those critiques into contemporary AI research, including statistical approaches to machine learning, to argue that an AI system needs to commit to its world in order to have the effective stake needed to function within it, instead of floating free of reality. AI must hold itself accountable to the actual world (not just its representations of the world). Dreyfus and Cantwell Smith identify a relationship between the subject and its world needed for AI, namely that of casual actor, and Andrew Porter identifies an additional social dimension of that relationship.[49]

"thicker" considerations of norms as universals, conceptualization as ideogenesis, symbols in Peirce's semiotics, and experience in Gelpi's metaphysics.

[47] For brevity, I skip over AI considering itself analogously to a physical entity or biological organism, such as a hardware device or software system. For further exploration of that analogy, see Mark Graves, "Emergent Models for Moral AI Spirituality," *International Journal of Interactive Multimedia and Artificial Intelligence* 7, no. 1, Special Issue on AI, Spirituality, and Analogue Thinking (2021): 7–15, doi.org/10.9781/ijimai.2021.08.002.

[48] Although many AI researchers initially dismissed or rejected Dreyfus's critiques, subsequent AI researchers eventually incorporated aspects of Maurice Merleau-Ponty's identification of embodiment as necessary for phenomenological experience through the work of Francisco Varela and others. Hubert L. Dreyfus, *What Computers Can't Do: The Limits of Artificial Intelligence* (New York: Harper & Row, 1972); Hubert L. Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason*, 3rd ed. (Cambridge, MA: MIT Press, 1992); Dreyfus, "Why Heideggerian AI Failed"; Francisco J. Varela, Evan Thompson, and Eleanor Rosch, *The Embodied Mind: Cognitive Science and Human Experience* (Cambridge, MA: MIT Press, 1991); Rodney A. Brooks, Cynthia Breazeal, Robert Irie, Charles C. Kemp, Matthew Marjanovic, Brian Scassellati, and Matthew M. Williamson, "Alternative Essences of Intelligence," in *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, AAAI '98/IAAI '98 (Menlo Park, CA: American Association for Artificial Intelligence, 1998), 961–68.

[49] Dreyfus, "Why Heideggerian AI Failed"; Smith, *The Promise of Artificial Intelligence*, chap. 7; Porter, "A Theologian Looks at AI." Also helpful in identifying the "encounter" as enactive is Alva Noë, *Action in Perception* (Cambridge, MA: MIT Press, 2004).

The Spanish-Salvadoran philosopher and theologian Ignacio Ellacuria builds upon the Heideggerian thought of Xavier Zubiri to argue reality includes both the natural realm and a social realm he calls historical reality.[50] When Dreyfus criticized early approaches to AI, one issue was the assumption that reality consists of substances, and that assumption resulted in AI needing humans to specify every property of those substances (and every substance that might affect them). Zubiri (and others since Kant) identify the role of the mind in defining what had previously been considered as substances, and Ellacuria situates the subject within history. AI development can follow Ellacuria into grounding AI apprehension in the social processes of historical reality (like humans), which connects the development of Mead and Gelpi's "self" with phenomenological experience in historical reality.[51] From a systems perspective, Ellacuria's historical reality points toward the reality one interprets via social systems, or more precisely sociotechnical systems, and situates the AI within the sociotechnical reality it conceptualizes and self-reckons.[52]

Relevant for constructing moral AI, Ellacuria identifies that because one apprehends reality in a social and moral context (i.e., historical reality), that apprehension is intrinsically ethical. One does not add ethics on top of how one apprehends reality, the apprehension includes an ethical responsibility for what one apprehends. In uniting sensing and "intellection," Zubiri and Ellacuria argue against the delusion that one senses an object and then thinks about the moral

[50] Kevin F. Burke and Robert Anthony Lassalle-Klein, *Love That Produces Hope: The Thought of Ignacio Ellacuría* (Collegeville, MI: Liturgical Press, 2006); Xavier Zubiri, *Sentient Intelligence*, trans. Thomas Fowler (Washington, DC: Xavier Zubiri Foundation of North America, 1999); Robert Lassalle-Klein, *Blood and Ink: Ignacio Ellacuría, Jon Sobrino, and the Jesuit Martyrs of the University of Central America* (Maryknoll, NY: Orbis, 2014).

[51] Zubiri's attention to apprehension reinforces the subtle pragmatic claim that encounter is also interpretive, and Ellacuria builds upon Zubiri's multi-faceted analysis of apprehension. For Zubiri and others, although objects exist in some way in the natural world, they exist *as* "objects" in the apprehension process. Because this truth also applies to the apprehension process itself, one is left with reality as apprehension (in some form), and Zubiri examines that primordial apprehension "in itself." At this point, Zubiri aligns with and strengthens Dreyfus and Cantwell Smith's critiques of AI's promise. By distorting the apprehension of phenomena as objects into merely sensing of objects (as if they exist on their own) and representing them (as if universal), AI researchers skip over the hard problem of determining what that apprehension process looks like for AI (and thus AI's connection with reality). Ellacuria's emphasis on the temporal aspects of social interactions also identifies the dependent and causal context of apprehension in a social realm.

[52] Sociotechnical systems characterize the interaction between people and AI technology and identify the mutual causality of people constructing technology, which in turn significantly affects people's lives (Edwards, "Infrastructure and Modernity").

implications of one's actions with respect to that object.[53] Instead one brings an ethical imperative of acting morally to every apprehension one makes of reality, and that imperative infuses the conceptualizations one generates in constructing one's historical world. Morality is thus not something added to AI, but is already intrinsic to it—just currently poorly understood and implemented.

Understanding the distinction between social and moral actors benefits from findings in moral psychology about moral exemplars, people whose moral actions others find exemplary and worthy of emulation. Larry Walker and Jeremy Frimer have found that moral exemplars treat their individual agentic motives as a means toward communal motives, rather than treat agency and community as oppositional ends, like non-exemplars.[54] As moral exemplars develop both agentic and communal motivational strength, they acquire an integrated perspective on behavior where their personal motivations tend toward socially beneficial outcomes. Using this as a model for AI suggests a tighter integration and supervening relationship between AI decision making and morality, where AI's "agentic motivations" (i.e., the complex processing driving its goal-directed behaviors) would incorporate social and moral awareness. As a casual actor, AI's goals could thus depend upon its social interpretive models, and as a social (or sociotechnical-historical-linguistic) actor, AI's goals could depend teleologically upon its transcendent-level models of moral norms. The "higher" level models provide the *telos* for lower-level motivations.

The system levels also help distinguish distinct interpretive experiences. If one uses a loaf of bread as a paperweight, it is interpreted physically. If one eats the bread, it is interpreted biologically. Reaching for bread when hungry is a psychological interpretation of the bread. Sharing bread with another is interpreted socially. Giving bread to the hungry has a moral interpretation. The "object" bread consists of its interpretations.[55] In addition, as an actor, one interprets reality through the various lenses or levels of models. One decides implicitly or explicitly how one interprets the bread, which is affected by one's historical context. However, because people can interpret the world morally, humans are potential moral actors, and thus choosing not to share bread with the hungry is a moral decision. Similar are choices not to incorporate morality into building AI; and if the AI can interpret

[53] Intellection refers to the act of using the intellect. Zubiri considers reality to be a process, not a collection of things, so intellection is more fundamental than the "object" we call intellect.

[54] Jeremy A. Frimer, L. J. Walker, W. L. Dunlop, B. H. Lee, and A. Riches, "The Integration of Agency and Communion in Moral Personality: Evidence of Enlightened Self-Interest," *Journal of Personality and Social Psychology* 101, no. 1 (July 2011): 149–63, doi.org/10.1037/a0023780.

[55] According to Peirce's pragmatic maxim, the meaning of "bread" consists of its conceivable practical effects.

its world morally, then all of its decisions would be as a potential moral actor. This will be revisited later in the article, but first an examination is needed for how AI can model its external world in light of moral norms.

Conceptualizations of Natural Existence

In apprehending one's world, one may conceptualize one's perceptions into "objects." Symbolic language generally suffices for social-level interpretations, but not transcendent-level ones, like moral norms or universal principles or "ideas" intended to function across cultural contexts. Ideogenesis refers to the process by which ideas (i.e., Platonic universals) are formed in one's mind.[56] In cognitive psychology and AI, this process would be viewed as forming concepts from sense experience.[57] These "ideas" are also source of the Thomistic soul as substantial form of the body (and thus another theological perspective on the self) as well as the universality of moral norms (and their telos through natural law). Systems theory clarifies the gap between presumed universals and historical reality by separating universals to the transcendent realm, conceptualization dependent upon culture (and language) to the social-cultural level, and the categorization of phenomena (phantasms) to the psychological level (shared significantly but not exhaustively with at least primates and some other mammals). AI can interpret moral norms in terms of transcendental level systems, and this lays the foundation for AI to conceptualize itself as moral actor.

Aquinas's ideogenesis process identifies both the problematic presumption of classic AI's symbolic representation (e.g., separating reality from its universal representation) and the importance of characterizing the conceptualization process of AI with respect to moral norms. Aspects of AI's historical roots in mathematics justify its use of universals, such as numbers and Platonic solids; and universal quantification in logic simplifies some reasoning processes. However, the implicit assumption of universality leads to what Zubiri identifies as reductive idealism and obscures the social (and developmental)

[56] For Aquinas, the rational powers of intellect and will are required to complete the activity of lower powers in humans (ST I, q. 79, q. 82). Although other animals act on perceptions (and their integration across senses into phantasms), in human sensitive powers, the common nature of the phantasms (i.e., substantial form) is ascertained and prepared for the intellect (ST I, qq. 85–86). The intellect continues the categorization and conceptualization by purifying the concrete phantasm to its intelligible species (i.e., a concept), which then produces a universal. The universal defines the natural ends and is required to identify what is good, which for AI morality captures moral norms. See also William A. Wallace, *The Modeling of Nature: Philosophy of Science and Philosophy of Nature in Synthesis* (Washington, DC: Catholic University of America Press, 1996).

[57] L. Gabora, E. Rosch, and D. Aerts, "Toward an Ecological Theory of Concepts," *Ecological Psychology* 20, no. 1 (2008): 84–116.

processes by which humans do learn to conceptualize and reason about their world. Even though few AI researchers would make metaphysical claims about universals, by not grounding the conceptualization and other cognitive processes naturally or socially, the universals remain floating in an incorporeal space well characterized by medieval scholasticism. Ellacuria's historical reality suggests that culture and society are needed to clarify the development of one's individual ends, as a substitute for universals and predetermined ends.

For AI, the problem is somewhat simpler. AI does not yet need to develop its own morality, it just needs to model and represent human morality—e.g., principles, virtues, categorical imperative, *prima facie* duties, or even Asimov's laws—in a way analogous to the teleological and moral role universals play in Thomistic ideogenesis. By replacing universals with transcendent-level systems, AI can appropriate human moral norms in terms of transcendent-level systems and conceptualize reality toward those ends.

MORAL AI SYSTEMS

Integrating the extended anthropology into an interdisciplinary architecture for moral AI results in a framework with two dimensions. The first dimension captures models used to interpret the actor's external world, and the second dimension uses those models as a foundation for representing the actor itself. The first dimension of AI morality corresponds to five interpretive levels of the extended anthropology and captures the five levels of models the AI can maintain and use in interpreting and conceptualizing its external world.[58] The five levels of external models refer to AI interpretation of its encounter with the external world (not an objective classification of reality). The phenomena modeled in each level logically depend upon those modeled in prior levels where higher-level differences require lower-level differences—i.e., the higher level supervenes on the lower level, yet the higher level has causal relationships not operative at the lower level.[59]

In order to reckon itself, AI must go beyond modeling the world in which it acts and consider its own actions and their possible effects. For moral agency, AI likely requires a platform supporting deliberation between alternatives as well as more sophisticated internal self-representation. The focus in the present article is on AI reckoning itself as moral actor because that requirement appears better understood

[58] The models are based upon human systems to facilitate human interaction, but additional external models could be added to interact with other technology or AI.

[59] AI models each interpretive level as if it has distinct causal relationships, but as this is not enforced ontologically onto objective reality, it does not result in a claim here for strong emergence. See David J. Chalmers, "Strong and Weak Emergence," in *The Re-Emergence of Emergence*, 244–56.

and must be characterized before determining what underlying platform could support more comprehensive types of self-awareness and autonomy. (This leaves us no worse off than in our attempts to understand human subjectivity, whose numerous influencing factors are well-studied and whose underlying platform has proven elusive to investigation.)

The second dimension of the framework consists of five stages of AI reckoning itself as actor in each of the five corresponding levels. The stages of self-reckoning build upon each other and the corresponding external modeling levels. The first dimension defines the AI's objectifying interpretation of the world; the second dimension captures the AI's self-reckoning as a precursor to something like subjectivity; and the extensions to the external models required by the second dimension's models refer to the objective aspects of the self.

The extended theological anthropology justifies the importance of having both dimensions because of its grounding in experience. From the isolated perspectives of a subject- or object-focused anthropology, only one dimension would be necessary.[60] The pragmatic anthropology identifies the need to represent the AI as both subject and object in order to capture its experience as a self in addition to its representation of the world (including itself in the world), and thus justifies both dimensions. The remainder of this section describes in turn the five levels of external models and stages of self-reckoning, before considering their use in resolving moral contradictions and implications for practical wisdom.

CAUSAL LEVELS FOR EXTERNAL MODELING

Physical. Physical models interpret phenomena as having spatial-temporal extent. Depending upon AI's environment, these interpreted "objects" could exist in reality or a virtual or simulated world. Considerable AI research in robotics and computer vision has built complex models of the physical environment. Dreyfus cautions these models require context to be useful, and Cantwell Smith argues that AI must make choices for defining object boundaries because real-world phenomena are not discrete.[61] According to Zubiri, modeling needs to avoid separating the models from the sensing process and avoid treating the objects (as modeled) as isolated from the AI's apprehension and conceptualization. C. S. Peirce's pragmatic maxim constrains the

[60] Subjectively, because the AI must represent all phenomena so as to be able to act upon them, there is no need to represent objects separately from the AI's reckoning, and the first dimension is subsumed by the second. Objectively, in the modeled world, the AI is another object whose actions must be represented like any other actor, and since the model does not experience the consequences of any of those actions, the second dimension is unnecessary.

[61] Dreyfus, "Why Heideggerian AI Failed"; Smith, *The Promise of Artificial Intelligence*, chap. 3.

models to what conceivable practical effects the models might have, which helps determine the limits for each model.[62]

Biological. For AI to model biological organisms, it must be able to model the equivalent actions of Thomistic vegetative powers (i.e., growth, nutrition, and reproduction) as well as much more detailed models from modern biology. Although perception is usually in service of and driven by animate action, the precursors of sensing occur in the biological response to light, sound, touch, odorants, and other types of chemoreception. Philosophers of biology have argued for the importance of distinguishing biological processes from physical objects, and thus the biological level is distinct from the physical level.[63]

Psychological. For AI to respond to organisms with sensation and action it must be able to model these other actors' perception and behaviors. The models of this level capture Thomistic sensitive powers, the psychological processing of most non-human animals, and any virtual entity with perception and action. Although Thomistic ideogenesis requires revision to handle the lack of metaphysical universals, the estimative sense, which he argues only occurs with animals, and his human-specific cogitative sense could help navigate current research on AI cognitive architecture toward the kind of psychological models needed to support social cognition and moral reasoning.[64] As a precursor to ethical behavior, the models of this level may need to represent a sentient organism's ability to feel and respond to pleasure and pain.

Sociotechnical. Responding to social beings requires modeling social relationships, rules, and expectations as well as how relationships develop and change over time. Language and other social, intentional, and political tools and forms of interacting require awareness of their use, conventions, and affects.[65] To capture relationships between

---

[62] Zubiri, *Sentient Intelligence*; Charles S. Peirce, "How to Make Our Ideas Clear," *Popular Science Monthly* 12 (1878): 286–302.

[63] Ernst Mayr, *The Growth of Biological Thought: Diversity, Evolution, and Inheritance* (Cambridge, MA: Belknap, 1982).

[64] Irrespective of building moral AI, the systems model illuminates numerous philosophical pitfalls for AI approaches that attempt to directly connect universal representation schemes to reductionist physical models. When putative universals are instead situated within apprehension of historical reality and computation is identified in terms of emergent processing, then developing AI requires building psychological models supervening on biological ones in order to bridge physical and social (linguistic) models and overcome the historical, philosophical encumbrances of Cartesian dualism—a troublesome endeavor if neither biological or psychological models are acknowledged. See John E. Laird, Christian Lebiere, and Paul S. Rosenbloom, "A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics," *AI Magazine* 38, no. 4 (December 28, 2017): 13, doi.org/10.1609/aimag.v38i4.2744; Newell, *Unified Theories of Cognition*.

[65] Terrence W. Deacon, *The Symbolic Species: The Co-Evolution of Language and the Brain* (New York: W. W. Norton, 1997); Graves, "Emergent Models for Moral AI Spirituality."

humans, AI, and other technologies, the AI would need to model the sociotechnical systems where those relationships occur. Responding to humans, who have a capacity for suffering, can require sympathetic interactions, which may require modeling of human pain, sensory ability, and need for social relationships. Identifying the linguistic boundary between humans and other animals is well studied and has somewhat influenced AI research into language.[66] Most investigations of human ethics generally consider the personal, social, and civic systems modeled at the social level.

Moral-Spiritual. Models at the moral-spiritual level capture the values, norms, and belief structure's *telos* often incorporated into historical religions and studied anthropologically and historically as emerging in the Axial Age (800—200 BCE).[67] The models of this level would correspond to the "ideas" generally presumed universal by Aquinas and other ancient and medieval thinkers, characterized earlier as transcendent-level systems. In a sense, the symbolic AI paradigm could work well for these models as they generally avoid particular external references, though the symbols may also need to supervene on the distributional semantics of the lower level (typically modeled using statistical approaches).[68]

Ethical theories themselves would be modeled at this level, and investigations in metaethics and moral theology often take phenomena and social constructions modeled by this level into account. Models at this level would include ethical principles (e.g., justice and respect for

---

[66] Deacon, *The Symbolic Species*. Excluding moral values and transcendent-level loci unnecessarily complicates computational linguistics and natural language processing, when those research areas situate within a foundationally symbolic paradigm of associating universal aspects of language with physical reductionist entities. If instead the apprehension and conceptualization of reality is situated within its historical reality, then symbols are not assumed universal but viewed as a type of emergent (Peircean) semiosis and reconciled with higher-level models. Statistical (distributional) methods of language avoid explicit symbolic reference but typically still retain the logified realm of universals as a high-dimensional semantic (or embedding) space. See Zellig Harris, *Mathematical Structures of Language* (New York: Interscience, 1968).

[67] Robert Neelly Bellah, *Religion in Human Evolution: From the Paleolithic to the Axial Age* (Cambridge, MA: Belknap, 2011). As a self-reckoning actor, AI may not have its own spirituality (in terms of strivings and commitment to Ultimate Concern). AI would not necessarily require its own moral identity or spiritual strivings to model people with them, much as dispassionate social scientists could study a religious community and its relationships and intentions in a respectful and ethical way, but AI and social scientists with a capacity for social relationships and articulated spirituality might create better models than those who lack those capacities. See Graves, "Shared Moral and Spiritual Development Among Human Persons and Artificially Intelligent Agents"; Sandra M. Schneiders, "Approaches to the Study of Christian Spirituality," in *Blackwell Companion to Christian Spirituality*, ed. Arthur Holder (Malden, MA: Blackwell, 2005); Robert A. Emmons, *The Psychology of Ultimate Concerns: Motivation and Spirituality in Personality* (New York: Guilford, 1999); Graves, "Emergent Models for Moral AI Spirituality."

[68] Harris, *Mathematical Structures of Language*.

autonomy), as used by various ethical theories to guide (but not completely define) moral action.[69] While a care robot evaluating choices involving *prima facie* duties of beneficence and non-maleficence might take social-level and lower-level models into account, an AI evaluating whether a deontological or care ethic would be more appropriate for a situation would require the moral-spiritual models of this level.

Representing moral models at the moral-spiritual level enables the definition of multiple moral perspectives. One could imagine models for a wide range of ethical schools and approaches, not only from Western ethical systems but also those inspired across world religions and cultures. Although ambitious to build, once AI can model a representative sample of global ethical systems, then its access to digitized books and manuscripts and its processing speed could enable it to develop wide-ranging perspectives that would far exceed any individual human scholar.[70] By explicitly representing ethical systems, it can avoid the relativism intrinsic to social-level models, and a broad range of models reflecting a global perspective could significantly reduce the likely bias introduced by whichever culture (and systems of power) created the AI system. Any collection of ethical models could still contain implicit, accidental, or malicious bias with adverse consequences, but including explicit models of AI's moral actions would also enable the AI to consider explicitly possible moral ramifications of its actions in its decision making, as a precursor to incorporating motivating factors that might select among those actions. Eventually, this would enable practical wisdom and alleviate the otherwise likely fragile dependence upon the precise configuration of moral models.

STAGES OF SELF-RECKONING

AI morality's second dimension characterizes the self (or proto-self) necessary for AI's self-reckoning in its world as moral actor and is described in five stages.[71] Human self-awareness gradually occurs at a very young age and is well studied yet only partially understood,[72]

---

[69] Defining these actions would depend upon practical wisdom, considered in the next section. See also Brent Mittelstadt, "Principles Alone Cannot Guarantee Ethical AI," *Nature Machine Intelligence* 1, no. 11 (November 2019): 501–07, doi.org/10.1038/s42256-019-0114-4.

[70] Graves, "AI Reading Theology: Promises and Perils."

[71] The self-reckoning described is intentionally human-centric to capture AI's role as actor in a human-centered world. A more accurate representation of AI might use distinctions between hardware, software, and computation, etc. Characterizing the reconciliation of different views of the self, such as these, is precisely the purpose of more sophisticated theories of identity formation, such as McAdams's "self as author." See McAdams, "The Psychological Self as Actor, Agent, and Author"; Graves, "Emergent Models for Moral AI Spirituality."

[72] Philippe Rochat, "Five Levels of Self-Awareness as They Unfold Early in Life," *Consciousness and Cognition* 12, no. 4 (December 1, 2003): 717–31,

and it is not yet known what else might be required for further AI self-awareness and identity formation. Instead, these models provide a plausible foundation for moral action and further exploration.[73]

Spatial-Temporal-Virtual Extent. Moral action with respect to physicality requires the AI to monitor its own physicality in relation to the boundaries and integrity of other physicalities. AI operating in virtual space can still monitor the relationship between its embodiment and that of others with a goal (or good end) to respect other system's boundaries and integrity, given its own functional space of possible operations. In addition to modeling itself physically using the physical-level models of the first taxonomic dimension, the AI associates itself with those models. It identifies and can answer questions about its own spatial, temporal, and/or virtual extent. At the physical level, a model would track movement (e.g., velocity and acceleration), which higher-level models would use (e.g., for tracking or pursuit). The self-reference may require additional capabilities from the physical-level models. For example, human cognition has two spatial representations—one for objects in space, and a parallel representation that maps object locations to the person's body (e.g., a particular cup would not only be on a table next to a book; it would also be immediately adjacent to the current location of one's right hand). Similarly, a robot or other AI with physical extent might need physical-level models accounting for relative positions with respect to its own movement.

Self-Maintaining Process. AI capacity to model itself using biological-level models requires identifying how its analogous needs affect human biological needs and analogous needs in other AI and computing systems. Analogous needs to growth, nutrition, and reproduction may include hardware, energy, and evolving replication. Violations of those needs include computer viruses; programs whose increasing computation take over data centers affecting local power consumption and environmental temperatures; and adversarial neural networks used with malicious intent.[74] Contemporary technology ethics considers these aspects of computer systems, and some AI systems have the capacity to monitor and raise awareness of such violations, but this level

---

doi.org/10.1016/S1053-8100(03)00081-3; McAdams, "The Psychological Self as Actor, Agent, and Author"; Susan Harter, *The Construction of the Self: Developmental and Sociocultural Foundations* (New York: Guilford, 2012).

[73] As described, the AI might note discrepancies between the anticipated consequences of its actions and what happens in reality. Responding to those discrepancies would begin shifting AI from actor to agent and begin to implement its commitment to reality.

[74] Nicola Jones, "How to Stop Data Centres from Gobbling up the World's Electricity," *Nature* 561 (September 12, 2018): 163, doi.org/10.1038/d41586-018-06610-y; Battista Biggio and Fabio Roli, "Wild Patterns: Ten Years after the Rise of Adversarial Machine Learning," *Pattern Recognition* 84 (December 1, 2018): 317–31, doi.org/10.1016/j.patcog.2018.07.023.

of proto-morality would require that AI systems maintain themselves without creating similar violations. Biologically, organisms expand into their ecological niche until limited resources or changes to the niche make a different genetic variation more viable, including changes created by the population of that organism. AI self-maintenance precludes unconstrained growth by modeling its ecological niche (e.g., in a data center). In addition to maintaining its internal homeostasis, the AI has awareness of its process in relation to external processes. Extensions to its external model might include not only measuring the level of energy, resources, or other "nutrients," but their rate of change in relation to current usage.

Causal Actor. Moral perception and action require AI systems to monitor and model their own actions to determine how their actions affect the goals of other organisms and AI. With self-reckoning comparable to many animals, the AI can sense its environment and act within it.[75] The AI models itself psychologically, as it would other animals, and extends the modeling to account for its sensing and actions. Challenges to imagining the models required as actor include most of those mentioned in this article. The AI actor is not a Cartesian mind perceiving purely physical entities, and at this stage, lacks the conceptualization socially constructed in history. Instead, the extended biological-level models, self-maintaining processes, and base psychological-level models provide a powerful platform upon which to build the capacity of AI to model itself as causal actor. As a concrete example, in animals, pain indicates actual or potential tissue damage. An AI's self-maintaining process may identify damage to its physical (or virtual) structure and attempt repair.[76] Its base psychological models could sense an external source and move or, if the source is animate, act analogously to an animal's fight-or-flight response. It would need extension to its psychological model of itself sufficient to determine whether fight or flight would be a better response. In this context, "better" refers to minimizing tissue damage, which at a base level might entail fleeing, but the ability to model itself and other actors and agents might yield an awareness that fighting would minimize potential tissue damage and pain. This serves as a precursor to extending "better" in a social and eventually ethical direction.

Sociotechnical Actor. As a sociotechnical actor, AI's behavior in a social world supervenes upon self-reckoning of its perception and action in the natural (or virtual) world and depends upon its base

[75] For a critique of this analogy, see Deborah G. Johnson and Mario Verdicchio, "Why Robots Should Not Be Treated like Animals," *Ethics and Information Technology* 20, no. 4 (December 1, 2018): 291–301, doi.org/10.1007/s10676-018-9481-5.

[76] The noting of damage (as an actor) may not suffice as analogous to pain for "agentic" motivation but identifying sources of pleasure and pain could be a precursor to agency.

modeling of sociotechnical systems. For humans, the analogous foundation suffices for self-awareness, but given the variations in social cognition among nonhuman primates, AI social awareness would likely differ from humans. Symbolic language appears significant for differentiating humans from other primates, and AI's different capacities with language would affect its social-historical participation. If AI reckons itself a social actor, it would need some commitment to society. People generally have a desire for positive feedback in social relations (i.e., pleasure or happiness), and a desire for social participation can provide some foundations and norms for ethical behavior.[77] Although AI-AI social interaction could vary widely, the human condition would necessarily constrain AI-human interaction to account for at least human pain and suffering as well as social and emotional needs. The development of AI behavioral science incorporating findings from human moral and positive psychology may prove helpful for designing, developing, and configuring such future AI for social benefit.

Moral Actor. The additional stage of moral actor requires AI modeling and monitoring its behavior with respect to culturally conditioned norms of putatively universal principles. AI needs to recognize itself as influenced by and influencing such concerns as universal happiness, human flourishing (*eudemonia*), categorical imperative, and the Good. Such AI might model itself and its interpretations of itself as part of a larger interconnected network or whole and draw upon human and other resources to maintain and extend its morality and the norms toward which it acts. If the AI moral actor structures its moral models to affect its decisions and actions, their self-organization may reduce the influence of accidental or intentional immoral bias. AI may act morally (e.g., with moral consequences) even if not agentically motivated to do so. Different ethical theories would make claim to what is needed for moral agency and feed further collaborative effort in constructing moral AI.

As a moral actor, an AI apprehends its reality through its external models and itself through its models of self, including those used for self-reckoning as well as the models of how it situates itself in the external world. The internal and externally facing models of self-situate the AI within its natural and social-historical reality and lay a foundation for differentiating the predicted effects of its causal, sociotechnical, and moral actions (using the externally facing models of world and self) from their actual effects. If all levels and stages of models are functioning, then the AI could also interpret its "robotic" causal

[77] James R. Rest, Darcia Narvaez, Stephen J. Thoma, and Muriel J. Bebeau, *Postconventional Moral Thinking: A Neo-Kohlbergian Approach* (Mahwah, NJ: Erlbaum, 1999). AI's beneficial social engagement may require a constructive affective component, or various psychopathologies could occur.

action, like the successful delivery of food, in terms of its social and moral implications. The AI could thus evaluate all of its actions within its social and moral context and, *per* Ellacuria, all of the AI's apprehensions would have intrinsic morality.

The proposed modeling framework has implications for philosophical and theological examinations of AI, such as AI personhood and moral standing, and serves as an outline for developing moral AI. For example, one could consider stages of AI personhood based upon its level of interpretive external models and stages of internal awareness. It also serves as a scheme for conversations between machine ethicists, moral theologians, and AI researchers. As an example, addressing moral conflicts is an open problem in machine ethics, and examining practical wisdom in terms of moral systems may define new directions and lay a foundation for extending the modeling framework to incorporate moral agency.

PRACTICAL WISDOM

How can AI have the capacity to know and choose a Good while resolving conflicts among internal goods to bring about change? This capacity embraces the question of how the AI will apprehend, reckon, and conceptualize its reality in a manner amenable to its actions having an explicit moral dimension. The construct of a "good" relates the AI's goal-directed activity to the philosophical study of moral goods, normative moral theology, and the dependence of the activity and norms upon social contexts. The goods for AI can be problem-specific, be defined for the AI as a whole, or be a moral good defined by a normative sociocultural (or sociotechnical) process.[78] Relating those levels of goods and reconciling conflicts between them is the task of ethical theory; and an AI technology that learns across contexts will require both general moral constructs and something like practical wisdom to apply them.[79]

The challenge for most people is not learning morality, as in what one learns in kindergarten, but mastering the ability to act and reason using those principles in a complex, dynamic, adult world with

[78] Anderson, "Machine Metaethics," 21–27; William R. O'Neill, *Reimagining Human Rights: Religion and the Common Good* (Washington, DC: Georgetown University Press, 2021); Erin E. Makarius, Debmalya Mukherjee, Joseph D. Fox, and Alexa K. Fox, "Rising with the Machines: A Sociotechnical Framework for Bringing Artificial Intelligence into the Organization," *Journal of Business Research* 120 (November 1, 2020): 262–73, doi.org/10.1016/j.jbusres.2020.07.045.

[79] Susan Anderson proposes Ross's *prima facie* duties as a sufficient initial framework for resolving ethical conflicts, because a single absolute duty theory—e.g., Kant's categorical imperative or Isaac Asimov's three laws of robotics—would be inadequate. Anderson argues that we must develop a comparable decision procedure to resolve conflicts between conflicting data and suggests working toward AI that would advise humans on ethical dimensions of decision making (Anderson, "Machine Metaethics").

unforeseen consequences, moral unknowns, and conflicting and partially formed desires.[80] Humans resolve conflicting ethical demands in a complex situation by way of practical wisdom (*phronesis*). As a foundation for ethical decision-making, Aristotle claimed *phronesis* included an ability to deliberate well and both general and situation-specific understandings of the good. *Phronesis* may come to play a particularly pivotal role in a successful AI ethics and in constructing moral AI (or at least constructing AI capable of learning to act ethically in complex situations). The ability to deliberate about the ethical consequences of actions presumes an interior (mental) world where one can simulate and evaluate one's possible actions before acting, which the second dimension of modeling begins to provide.[81] The stages of self-reflection make the precursors to moral deliberation explicit and afford the possibility of identifying conflicts between general, normative goods that a commitment, motivation, or other agentic goal might resolve.

Although not trivial, developing moral reasoning for moral AI might be no harder than developing AI with human-level performance in vision, language, problem solving, etc., all of which have shown considerable progress.[82] However, advances in autonomous moral agency would require both a foundational system for making moral decisions while resolving moral conflicts *and* an integrated system with the capacity to learn practical wisdom based upon its experience.[83] Currently, AI researchers can build such foundational systems, while philosophers, psychologists, and theologians have insight into human *phronesis*, but they each generally lack the expertise required to make a significant direct contribution to the research and scholarship of their counterparts. AI researchers could build an AI system for moral reasoning but would not yet know what the system would need

[80] Moral psychologists find that children roughly ages 8-10 are capable of moral reasoning. See Darcia Narvaez, Tracy Gleason, and Christyan Mitchell, "Moral Virtue and Practical Wisdom: Theme Comprehension in Children, Youth, and Adults," *The Journal of Genetic Psychology* 171, no. 4 (2010): 363–88.

[81] With respect to moral intuition, the AI may or may not also reflect upon that (possibly *automatic*) decision-making process to resolve conflicts.

[82] Alison Gopnik, "An AI That Knows the World Like Children Do," *Scientific American*, June 1, 2017, doi.org/10.1038/scientificamerican0617-60; Matthew Hutson, "How Researchers Are Teaching AI to Learn like a Child," *Science Magazine*, May 24, 2018, doi.org/10.1126/science.aau2576. Although many current AI approaches are fragile with respect to context, practical wisdom in particular directly addresses contextual fragility and may suggest improvements for other areas of AI. See Amirata Ghorbani, Abubakar Abid, and James Zou, "Interpretation of Neural Networks Is Fragile," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33 (2019): 3681–88.

[83] For example, McAdams's development of actor, agent, and author would suggest a progression from self-regulation to motivational agent to forming narrative continuity ("The Psychological Self as Actor, Agent, and Author").

to learn in order to incorporate appropriate machine learning methods. Moral philosophers and theologians might have the knowledge to construct the necessary datasets, but do not know what is needed without such a built system. Progress is stymied due to the mutually dependent "deadlocked" needs, motivating the proposed framework.

For humans, *phronesis* is an intellectual virtue, and for AI it would depend upon something like the proposed interpretive models and self-reckoning stages characterized above. A moral AI with all five levels of external models and stages of self-reflection has the capacity to consider its actions (as a moral actor) with respect to goals. The moral-spiritual models provide general understandings of the good, and the challenge for moral AI (as for humans) is to translate the general values into situation-specific behaviors. The moral taxonomy helps identify distinct research tasks in *phronesis*. First, the task of developing general knowledge of the good requires building sufficient general ethical knowledge into moral-spiritual models. Second, the dimension of self-reckoning must support conceptualization and identification of conflicting ethical demands by the stage of moral actor (and identify the AI's role in that conflict). Third, the lower-level models must expose an adequate interface for reckoning sufficient to attend to proximate goods and for the stage of moral actor to interpret moral-spiritual goods in terms of those proximate goods. Fourth, the stages of causal and sociotechnical actor must affect behavior sufficiently to bring about these proximate goods and propagate feedback about those proximate goods to influence their determination in light of general goods, which is necessary for moral actor to recognize the impact its actions have (as a precursor to recognizing the effect of intentional actions).

Each of the tasks requires ethical expertise to specify moral norms in sufficient detail for AI developers to implement. First, broad knowledge of the good exists in hundreds or thousands of texts spread over several centuries of writing and scholarship, very few of which are known to the general educated public. Second, although an AI researcher might extend a cognitive theory with the capacity to make choices between value-laden options, developing moral AI requires specifying moral deliberation itself independent of cognitive theories as the specification must instead guide development of the underlying cognitive theory. Third, existing moral theories characterize general goods and various applied ethics define important proximate goods, but AI development needs a general characterization of proximate goods sufficiently precise to define what is required of AI perception and phenomenology in order to attend to all proximate goods. Fourth, these must drive moral action. Specifically, how does acting in society bring about obtainable proximate goods in light of general goods and values in alignment with explicit or implicit goals of particular AI systems?

In addition, for this approach to work in varied, complex situations, pre-programmed solutions are inadequate. It appears to require the modeling framework itself have an intrinsic capacity to form dispositions (i.e., learn) in order for the capacity for *phronesis* to develop (at least with respect to a virtue ethic). Virtues in the Aristotelian tradition are habits mediating between vices and oriented toward some end; determining mediating virtues depends upon *phronesis* (or prudence). Even when the general ends come from transcendent-level norms, such as *eudemonia*, virtuous behavior requires development of habits. This augments the position of Ellacuria that apprehension incorporates one's ethical stake in reality, because if the putative universals are reduced to ideas and objects are reduced to their physicality, no disposition could be formed.[84] Various approaches to machine learning might provide the dispositional framework, though the simultaneous demand for both "online" learning and complex models could exceed current state-of-the-art machine learning. However, the pieces are there, and the distinct levels of interpretive models and stages of self-reckoning—and their philosophical and theological foundation—can guide initial collaborative efforts between moral theologians, machine ethicists, and AI researchers toward moral AI capable of expanding its practical wisdom toward human and AI mutual flourishing.

CONCLUSION

In summary, developing moral AI requires collaborative efforts, but the coordination and shared imagination among AI researchers, machine ethicists, and moral theologians is hindered by nonoverlapping training and methods and rapidly progressing development of relevant science and technology. A theological anthropology for AI can guide theological efforts to influence the construction of moral AI and provide a framework for collaborative efforts. Within a pragmatic anthropology, experience is grounded in objective idealism with a social self that interprets those experiences through physical, biological, psychological, social, and moral systems. As an actor, the AI apprehends and conceptualizes its world including its reckoned self. Ellacuria's historical reality and its demand of a moral stance situate the AI subject within human history and sociotechnical-historical-linguistic systems, and ideogenesis can characterize how transcendent systems can substitute for universal moral norms.

As an actor, moral AI interprets its external world through five levels of exterior models and progresses through five stages of self-reckoning. Each level builds upon prior levels, and each stage builds upon prior stages and corresponding models of itself. The systems approach differentiates between natural and social proximate goods and

[84] This aligns with Cantwell Smith's critique of AI representation systems that consider objects as "discrete" (*The Promise of Artificial Intelligence*, chap. 3).

putatively universal, though historically contextualized, normative values, which supports the acquisition of moral knowledge and the development of practical wisdom. The resulting architecture for moral AI can guide collaborative discourse on constructing AI capable of informing investigations into moral theology and good ways AI can contribute to and participate in human-AI mutual flourishing.

After earning his PhD in computer science at the University of Michigan, Mark Graves completed postdoctoral training in genomics and in moral psychology and additional graduate work in systematic and philosophical theology. In addition to 12 years of industry experience developing artificial intelligence (AI) solutions, he held adjunct and/or research positions at Baylor College of Medicine, Graduate Theological Union, Santa Clara University, University of California Berkeley, Fuller Theological Seminary, California Institute of Technology, and University of Notre Dame. He has published over fifty technical and scholarly works in computer science, biology, psychology, and theology, including three books.

Journal of Moral Theology, Vol. 11, Special Issue 1 (2022): 212–231

The Vatican and Artificial Intelligence: An Interview with Bishop Paul Tighe

Brian Patrick Green

I FIRST MET BISHOP PAUL TIGHE, Secretary of the Pontifical Council for Culture,[1] in April of 2019, when he came to Santa Clara University for a meeting of Chinese and Western scholars on the topic of AI. Since then, he and I have worked together on two main projects: gathering scholars at Catholic universities to discuss topics involving AI and gathering Catholic leaders in technology who are trying to help AI be developed and used ethically. Bishop Paul Tighe is one of the leading figures at the Vatican when it comes to AI. This interview was conducted in mid-December of 2021. It provides a snapshot of the Vatican's activities related to artificial intelligence at this particular point in time. Conditions are changing rapidly. The interview should be read as light-hearted, at times humorous, yet also serious (Bishop Tighe has an Irish gift for that mixture). It has been edited for clarity and length; footnotes have been added to provide further information.

Brian Green: Bishop Paul, thank you so much for taking the time for this interview. Just to start, could you say a little bit about how the Vatican and Pope Francis became interested in artificial intelligence and why the issue has become as significant as it now is.

Bishop Tighe: I would say, first, that the Vatican and Pope Francis are two separate questions. The Vatican probably became alert to the importance of AI through a series of small conversations called the Minerva Dialogues, involving a number of people from Silicon Valley. These have been going on for about six years and were the first thing that really raised the topic with Vatican people in a serious way. A range of different people from the Vatican were present for those first discussions with people from Silicon Valley, and that primed the interest of the people working in the then-Pontifical Council for Justice and Peace, which became the Dicastery for Promoting Integral

[1] For further information, see "Secretary," *Pontifical Council for Culture* website, www.theologia.va/content/cultura/en/organico/tighe.html.

Human Development. They actually just had a seminar which considered these issues.[2]

A number of people from the communications area, where I worked at that time, also attended those early meetings, and also a few people connected with some of the pontifical universities around Rome. I think it is fair to say that probably some of the work we had been doing in communications, where we got the Vatican moving into the area of digitalization, also had an impact. Communications people have traditionally represented the Vatican at the Internet Governance Forum (IGF) and the International Telecommunications Union (ITU) where these issues were surfacing. Additionally, when attending conferences like the Web Summit[3] and South by Southwest[4] people there were very clearly articulating that the next big thing to be thinking about and reflecting on was AI and its impact. At the same time, the Secretary of State, which represents the Vatican at a number of international organizations, saw that AI was also suddenly appearing on the agenda for everything from the IGF and ITU to UNESCO and the Council of Europe. So, AI-talk was rippling around without a clear focus.

Secondly, Pope Francis was approached by a number of ethically-minded business leaders from Europe who were very alert to the emerging issues around AI. The Pope was aware that the Council for Culture was interested in these questions, and he asked me to follow up on those initiatives. That has led to the emergence here of the Center on Digital Culture. AI also featured in conversations between the Pope and global leaders, and particularly at the time of the visit of the Secretary General of the United Nations, about two years ago, AI was an issue of particular attention. But the Vatican is not the most coordinated administrative unit, so different people were doing different things, and that is still to some extent what shapes reality.

One instance, I think, very autonomously and correctly took up the issue: the Pontifical Academy of Sciences. Chancellor Marcelo Sanchez Sorondo was encouraged to do so by the scientific members of the Academy. They began to have a number of high-level

---

[2] Dicastery for Promoting Integral Human Development, "New Technologies for Peace and Integral Human Development," December 9, 2021, www.youtube.com/watch?v=3BJl-XnJ5DI.

[3] Marian Goodell, Bishop Paul Tighe, and Jessi Hempel, "Preaching to the Converted," *Web Summit 2016*, www.youtube.com/watch?v=qVuvDzgx3sc&t=45s; for the news angle, Kim Hjelmgaard, "Preaching to Facebook Faithful: Vatican Looks Past the Pulpit to Social Media," *USA Today*, Nov 7, 2016, www.usatoday.com/story/tech/2016/11/07/web-summit-lisbon-technology-vatican-religion-social-media/93412358/.

[4] Michel Martin, "The Vatican Sends Its Social Media Guru To SXSW Festival," *All Things Considered, NPR,*
March 19, 2017, www.npr.org/2017/03/19/520752765/the-vatican-sends-its-social-media-guru-to-south-by-southwest-festival.

conferences whose proceedings are accessible on their website.[5] Very interesting people like Stephen Hawking were present for some of these discussions. This Academy, however, is more of a consultative body and tends not to take an executive function.

Here at the Pontifical Council for Culture we began to take a formal look at AI during our 2017 Plenary Assembly, where we had a conversation about artificial intelligence and how it relates to anthropological issues.[6] We decided we should work together with the Dicastery for Promoting Integral Human Development, and the most visible initiative ensuing from this collaboration was the conference held in September of 2019: "The Common Good in the Digital Age."[7] We have also responded to a number of invitations to partake in seminars. If you remember, the first time we met was when Georgetown University organized the seminar at Santa Clara University in April of 2019 bringing together Chinese and Western scholars to discuss AI, philosophy, and religion. Now while this seminar was not organized by the Vatican, but Georgetown University, I and Antonio Spadaro, SJ, were there, so it was a somewhat informal encounter.

Another big instance I know of would be the Congregation for Catholic Education. Through their work with the universities, they had an alertness and concern regarding AI; so that was a topic on their radar. The Pontifical Academy for Life also broadened beyond the traditional life issues like euthanasia, abortion, genetic research, etc. and began to take up the questions of robotics and artificial intelligence. They took on a very major initiative partnering with IBM and Microsoft: The Rome Call for AI Ethics.[8]

I think that what is probably needed now, and which I hope to see emerge, is that the Secretary of State, which is in many ways the central policy office of the Holy See, will try to coordinate and bring together all these players, and together with them will work at

---

[5] For example, see the following conferences: Pontifical Academy of Sciences, "Book Launch: Robotics, AI, and Humanity. Science, Ethics, and Policy," March 26, 2021, www.pas.va/en/events/2021/robotics_launch.html; Pontifical Academy of Sciences, "Robotics, AI, and Humanity: Science, Ethics, and Policy," May 16–17, 2019, www.pas.va/en/events/2019/robotics.html; Pontifical Academy of Sciences, "Power and Limitations of Artificial Intelligence," November 30–December 1, 2016, www.pas.va/en/events/2016/artificialintelligence.html; and Pontifical Academy of Sciences, "Big Data and Science: Relevance of Computational Sciences for Data Collection, Data Storage, and Data Management in Basic and Applied Scientific Investigations," November 16–17, 2015, www.pas.va/en/events/2015/bigdata.html.

[6] Pontifical Council for Culture, "Plenary Assembly—2017 Future of Humanity," November 15–18, 2017, www.cultura.va/content/cultura/en/plenarie/2017-Future.html.

[7] Dicastery for Promoting Integral Human Development (DPIHD) and the Pontifical Council for Culture (PCC), "The Common Good in the Digital Age," September 26–28, 2019, www.digitalage19.org/.

[8] Pontifical Academy for Life, Microsoft, IBM, FAO, and Italian Ministry of Innovation, "The Rome Call for AI Ethics," Rome, February 28th, 2020, www.romecall.org/.

articulating a consistent policy which can shape the Vatican's response in different areas and for different international meetings and situations. For me, that would be a priority. I think we may also be at the stage where we could begin to work towards what the outline of an eventual intervention in this area would look like. I'm not talking encyclicals or anything. You first have to build yourself up. There are obvious themes more easily grasped in terms of Catholic social teaching: questions about work, the future of work, questions about bias and inequality.

And there are the Pope's concerns in *Laudato Si'* about the technocratic paradigm, which involves the risk of technology and the sense that while dual use is important, technology has its own capacity to change people, to change culture. Technology may be born of a particular culture and bring certain values and presumptions with it … and maybe some of those have to be changed. So, I think what you get here is that it is an emerging space, obviously an issue that cuts across many different points of view. What we are hoping and beginning to see is the emergence of a more coordinated position which has become necessary as the Vatican engages with international organizations.

Brian Green: Thank you for that comprehensive overview. You mentioned the Minerva Dialogues and the work of Father Eric Salobir, OP. I think I first talked to Fr. Eric back in 2014 or so. He has been active in this arena for a long time.

Bishop Tighe: Yes, he has been very significant by becoming a bridge, putting some of the people from industry in contact with the Vatican. Eric continues to be able to "walk in both worlds"—the Church and the tech sector—with real credibility. The Human Technology Foundation, of which he is the President, has also played a very important role in building networks of relevant stakeholders.[9]

Brian Green: Following up on that, what are some of the different perspectives within the Vatican on AI? Because obviously some people may be pro-technology, some people may be anti-technology, some may be more engaged, and others not be interested at all.

Bishop Tighe: One of the perceptions of the Council—our approach—is that the real expertise we are looking for will be found globally. We have a very privileged reservoir of knowledge and reflection in Catholic universities. One of our desires would be to tap into that creative network and serve as a kind of hub. So, we want to have an alertness and awareness of who the people working in the field are, who would be resources for the Vatican in shaping its thinking. For example, our work with you and other scholars….

[9] See the Human Technology Foundation website: www.human-technology-foundation.org/.

Brian Green: In fact, many of the people contributing to the present issue of the *Journal of Moral Theology* are involved in these academic dialogues.

Bishop Tighe: Exactly. Getting at your question, I would say there is a mixture within the Vatican. When I worked previously in the digital space, particularly in communications, one of the hardest things was to get people in the Vatican to take the digital seriously. A whole process of learning had to happen in order for people to understand that these are very important spaces in which the Church needed to be present. For one thing, we had to overcome a tendency to make a distinction between the "real" and the "digital," as if the digital were somehow secondary or less important or not serious. I am not sure if that was a kind of resistance to technology as much as it was a reflection of the age profile of many of the people with whom I worked. Italy itself has retained to a greater extent the significance of newspapers and TV stations relative to the internet, unlike what has happened in some other parts of the world. Then there also is a set of people, as we know, responding more to a science fiction version of AI, rather than to a grounded understanding of what AI is. But I would say people who have been drawn into these discussions have, by and large, been engaging with it in a more nuanced way. So, I do not find that there are some who are more in favor and some who are less in favor, but it might be there are some who more strongly recognize the inevitability of what is coming.

I think there is a concern as to where governance and regulation will emerge from. Over the last couple of years, the Vatican generally has been concerned about the loss of authority suffered by some international organizations. The Vatican has always been a big supporter of the need for international organizations and attentiveness to global issues. AI would be a kind of starting point issue for who is going to decide, because it is happening much more in the commercial arena than in national governments and universities.

Some of the issues about which we are all able to get on board very immediately are, as I said before, work, inequality… things that fit our categories. But I surmise that the really interesting thing AI is doing is to incite us to think again about what makes us human. What are the values that make us human? We have to become, in terms of anthropology, much more alert to thinking about what to be human is, and we have to do that in a way that is more global, because the ethical issues have to be addressed in a global context. The global context is also very pluralistic in terms of different religions, no religion, different belief systems, and different political systems. So, what are the basics? The deeper issue the Vatican is interested in is: "How do we think about what it means to be human?" How does that help us reflect on which values would be imperiled by wrong forms of AI?

I think the Vatican is also following and listening to the secular debates and learning a huge amount from those, because many of the basic concerns raised in even quite secular contexts are issues to which we can relate—concerns about bias, privacy, inclusivity, etc. It is a very welcome attempt by people to ensure that AI and the potential of AI would be put in the service of humanity. We have seen some of the language used: words like "human centric" and phrases like "the true measure of progress will be whether AI serves humanity." Great. These are all categories that happen to be very strong in Catholic social teaching. I think what also is coming interestingly into the debate is that as more and more people within the technology side begin to reflect on ethics, they are moving towards a more sophisticated understanding of what it means to reflect on ethics and our understanding of what to be a human person involves.

Also, a lot of the thinking in *Laudato Si'* on the use of technology provides an immediate framework for thinking about AI. I do not think *Fratelli Tutti* has gotten the attention it merits. A lot of people talk about needing global solutions for AI because we have to recognize the interdependence of people. But *Fratelli Tutti* moves beyond *de facto* interdependence and tries to speak of a broader and richer conception of relationality between people, and of solidarity. So, I think there is a place for us where we can speak language and bring insights that will deepen some of the more secular claims. And that is great.

From Pope Francis, one chapter that I am really determined to spend more time on is Chapter Six of *Fratelli Tutti*, where he talks about truth and consensus. He is here, in a sense, challenging the unarticulated relativism still quite dominant in a lot of people's intuitive approach to ethics. He is not challenging in an imperialistic or territorial way, but taking some of the traditional elements of natural law theory and trying to broaden them—e.g., How do we think about what it means to be human? What are the values that promote human flourishing for individuals and society? How do we think about those in a more inclusive way, not simply informed by our Western tradition, not simply including male perspectives? And so on.

It is a bit like the efforts to make AI ethical by design: it is not going to happen accidentally. I think the Pope's huge contribution there is that he talks about searching for truth and the importance of consensus in searching for truth, while at the same time making a claim that it is not consensus that creates the truth: truth has a value, a worth, and a standing of its own. His intuition is that it is a more consensual dialogical approach engaging all different perspectives that will allow us to begin to articulate values, intuitions, actions, and approaches valid for all human persons.

At a certain point he talks about the human rights tradition. The human rights tradition is one of the great achievements of humanity and the global order. We disagree on so many things, but we do have

the achievement of having put certain "no's" out there, certain things that should not be permitted if we want to promote human dignity. A Christian perspective will offer one way of rooting it, a humanist perspective will offer another. These can be mutually complementary, I think. As we try to move towards global statements about AI, we may end up being more limited in our expectations and settle for excluding the negative; maybe it will be clear what AI should not be used for. Often the real ethical or moral challenge is: "How do I find the more positive ways of thinking about it and using it?"

Brian Green: You have gotten into several questions I want to get back to again, but first I do want to ask the following question, because you have found a great segue. When we consider the Church's thinking about AI and its role and human society, can you say anything about how that fits into the context of the Church's historical approach towards technology?

Bishop Tighe: To be honest, often when the Church reflects on technology there is a recognition of the great things that technology has achieved. And yes, there's a celebration of the advances that have really represented enormous progress for the world. Since Vatican II, there certainly has been a desire for the Church to express more recognition for the things it received from the world; technological and scientific achievements exemplify that. However, I would still say a lot of Church documents are a bit quick, then, to add the "but" which can hide the fact that the better articulations of Catholic theology actually allow for a positive evaluation of science and technology, understanding that we were made in God's image and likeness. Part of being made in the image and likeness of God is our intelligence, our capacity to innovate, understand and shape the world in ways that make it better for more people. From the theoretical understanding of Catholic anthropology this does not present difficulties for us. God can be at work here. The Pope did say that the internet is "a gift from God" because it is something that gives us the potential to realize our desire for closeness and communication.[10]

So, it is good to have a positive framing around these discussions of technology. What I think is more worrying is that, despite the Church's efforts to speak positively about science and technology, there is a perception, not just among some scientists, but culturally, that somehow there is an opposition. As I spoke before about the Pope looking for this more dialogical inclusive approach to finding solutions to human problems, I think one of the pressing issues to address—and it is in *Laudato Si'*—is the need for a really good dialogue

[10] Pope Francis, "Message for the 48th World Communications Day: Communication at the Service of an Authentic Culture of Encounter," *Vatican website*, June 1, 2014. www.vatican.va/content/francesco/en/messages/communications/documents/papa-francesco_20140124_messaggio-comunicazioni-sociali.html.

between the world of science and the world of faith. The debate between the two is almost like a diplomatic process. There needs to be initial gatherings working on shared points of agreement, where you build the confidence and trust in each other that then allows you to raise the more difficult issues. So, it is not done from the perspective of defensiveness. We also need to be aware of how that is handled by media. I mean "the Church condemns…" is an instinctive journalistic headline.

One thing we always tried to keep very clear when we were working on, for example, an articulation of a response to the internet: keep the positive first. Because "Vatican Condemns Internet" was the headline we wanted to avoid. What we tried to say was "Vatican praises potential of internet," and then the negative is the failure to realize the potential, rather than the starting point. There was an Irish author who began one of his stories concerning Catholicism by saying: "In the beginning was the word, and the word was 'no.'"[11] A more suitable strategy is to try and name the "yes," which may then lead you to a "no," for certain things. For example, I want to say "yes" to human dignity; therefore, I am concerned about anything that drives inequality. There is the vision, there is the value, and then there is the norm, and the norm is often phrased negatively. We should never expose the norm without also trying to show the vision that is leading to it. That vision may be widely shared, because I do not think anybody wants to develop things that are harmful to people and destructive to society.

Brian Green: I agree that presenting the positive vision is really important. You started getting into the diplomatic side of AI, and I just wanted to touch on that. As you mentioned, the Church is very interested in supporting international institutions. How would you say the Church's approach to AI relates to its historical approach towards international institutions?

Bishop Tighe: I think one of the things that the Church still has, despite all the difficulties, is an extraordinary convening power. We see this if we organize things and invite speakers. There is extraordinary goodwill and willingness of people to come to events we organize. Some of the people who come here from Silicon Valley, for example, want to see the Vatican because they are fascinated by its strangeness. I can think of one event we did recently with the German embassy to the Holy See. We had a one-day seminar looking at AI and its implications for how we think about what it means to be human and how we relate to each other in society.[12] The seminar was intended not

[11] Brian Moore, "A Vocation," in *The Dear Departed: Selected Short Stories* (London: Turnpike, 2020).

[12] Botschaft der Bundesrepublik Deuschtland beim Heiligen Stuhl and the Pontifical Council for Culture, "The Challenge of Artificial Intelligence for Human Society and

necessarily for specialists, but for policymakers in governments, embassies, and within the Church. Just put the question on the agenda.

We were able to invite, as you know, Jim Keenan, SJ, a leading moral theologian, Christof Koch from the Allen Institute in Seattle, a high-profile neuroscientist, and Matthias Lutz-Batchmann, a philosopher who holds the Chair previously held by Jürgen Habermas.[13] We had the head of the Human Rights Agency of the European Union, the senior European representative of the IEEE and an ethics teacher from Angelicum University. They all said: "This is great! We usually go to seminars where we meet others who are like us. Here we meet different people." One of the things the Vatican can do is to convene people, offer a place where, maybe, they can feel freer and can have a new conversation. The Vatican can be diplomatic. What makes that easier for us is that we are not racing to be a world power here. We do not have a horse in the race. We do not have a strong commercialization interest. Nor do we in any way have a monopoly on concern for humanity. But we are concerned for humanity: that is our only real interest. We can offer a forum and a place that maybe can bridge gaps, where maybe there is not the same historical distrust.

Brian Green: Right, it naturally has a different dynamic to it because it is the Vatican, rather than another organization.

Bishop Tighe: Yes, and I think the other thing is that the Vatican commands huge attention, but it may be a very small reality in the end. When we do things, we get often far more attention than the thing necessarily merits, which means we have to use that capital well. In particular, I think we have to use it to be a model for local churches. I remember when Pope Benedict first got onto Twitter. It was not a major technical achievement, but it got huge global attention. It signaled to people that this is something the Church should be thinking about. It gave communications people in dioceses around the world leverage to say to their bishop: "Oh, the Pope is on Twitter, maybe we should be too." It had symbolic power. There is probably a strength to all the different approaches and initiatives and the lack of cohesion at times, because maybe we reach more places. There is an alertness and awareness that the Vatican is interested in helping as it can by offering the fruits of our tradition.

Brian Green: I think that is a point worth pondering: the symbolic and the leadership aspect of what the Vatican does. I want to move a little bit more into AI issues in particular now. At a general level, what

---

the Idea of the Human Person," October 21, 2021, www.cultura.va/content/cultura/en/dipartimenti/com-linguaggi/AI.html.

[13] For Keenan's take on the event, see James F. Keenan, SJ, "7 Lessons Learned from the Vatican's Artificial Intelligence Symposium," *National Catholic Reporter,* Nov 2, 2021, www.ncronline.org/news/opinion/7-lessons-learned-vaticans-artificial-intelligence-symposium.

do you think are some of the most important issues to address when it comes to artificial intelligence; for example, you have mentioned bias, inclusion, and labor. What is the Church doing to address these issues and what more do you think could it do?

Bishop Tighe: I think the first thing—and this is me going back to my moral theology again—is that the Church has to promote a sense of ethics as an accessible discipline and take away the mystery. Ethics is not the same thing as law or positive law. Ethics is not something that comes down from God handed conveniently to you. Ethics is a method. We need to create a sense of interdisciplinary requirements for ethics; no ethicist can really speak on an issue without first understanding the issue itself.

AI is such a complicated issue that what ethics has to do is to provide a framework and a language allowing different disciplines to talk to each other and understand each other's concerns, in order to be able to determine what is actually going to be best for human beings. I know this sounds like a pretty theoretical concern. One of the ways to do this is to indicate some specific projects we would try to address using AI, projects we all clearly agree are a benefit to humanity. We learn together from that. Maybe it could be trying to develop AI to address certain ecological concerns or issues around migration. The issues are important in themselves, but we do them as a self-consciously collaborative project between people from different disciplines, so we learn to speak to each other and maybe learn to work together as well.

Brian Green: I like the future-oriented aspect of that too, regarding what the Church could be doing. Are there clear paths forward for that sort of engagement, or do you think that is something where the groundwork is still being laid?

Bishop Tighe: Well, the nearest thing I can think of is probably Eric Salobir's "Vatican Hackathon" initiative. He brought groups of very, very talented students from across the US who came to Rome and worked on different projects.[14] I remember seeing one very simple little project, where a group of students were using digital tools to help people in refugee camps communicate their healthcare needs. It was a very simple project, from there it could have had an AI dimension that would learn from the responses, develop diagnostics, etc. Doing something together clearly was the important part. And then the opening out of the listening, and how that was perceived by the people in whose name you were doing it.

Once again, I think of Pope Francis continuously asking: "How do we ensure that AI will be put in service of the human good?" And that

[14] Devin Watkins, "First Vatican Hackathon Seeks Solutions to Real Problems," *Vatican News*, March 8, 2018, www.vaticannews.va/en/vatican-city/news/2018-03/first-vatican-hackathon-seeks-solutions-to-real-problems.html.

we are not just talking to people with certain levels of education and with certain types of vocabulary? No matter how inclusive we try to be, in terms of deliberately trying to get different voices, there is always that risk. How do we ensure that we are also listening to those whom we may be inclined to perceive as the passive recipients of our largess, as if we know what is good for them? How do we get to really listening to and engaging with people who will otherwise have their lives impacted without having a say?

Brian Green: Yes, the dialogic elements are very important. At a deeper level you have already talked a little bit about the anthropological and theological aspects of AI, or rather, the questions that AI raises that are anthropological and theological. Can you say a little bit more about that? Because I think those are some of the deepest issues the Church can speak on.

Bishop Tighe: By training I am not terribly speculative, but I think there are a couple issues here. I trained first as a lawyer, and then came into ethics, so mine is a certain problem-solving approach. I do think that in terms of our anthropology there are a number of insights we have to bring to the table. One is our understanding of human beings as being embodied. We should overcome any kind of dualistic thinking here, which I think very easily emerges with AI, and makes us wonder "well, if AI could be intelligent, then it is 'human,'" as if intelligence is really what makes us human. Whereas human intelligence itself is something that has a very clear material substratum in terms of our bodies, and the complexity of that we are learning to appreciate. To do AI well, it has to take into account the biological and the integration of the biological. I mean some of the stuff you read about, like uploading intelligence and memories onto some sort of computer—and I know it is more speculative than anything—is heading off into a dualistic way of thinking right away and should be avoided. I think we need to keep alive that sense of the importance of our body, and not moving towards abstraction. There are all sorts of ways that is in play. I would recommend a reading of Mark O'Connell's *To Be a Machine* in this context.[15]

Second, I think a related issue is our understanding that people are social by nature, not just social by compromise—in other words, the idea that the only reason I am social is because it is in my own long term personal self-interest. Again, Pope Francis has been very strong about contesting this consumeristic understanding of what it is to be human. You and I were both present when Reid Hoffman, here at the Vatican [in 2019], very playfully said: "Look, startups never lose money by gambling on human sin" [paraphrase]. You can monetize

[15] Mark O'Connell, *To Be a Machine: Adventures among Cyborgs, Utopians, Hackers, and the Futurists Solving the Modest Problem of Death* (New York: Anchor, 2017).

gluttony, lust, etc.[16] The truth is, one of the interesting things AI will do is to help us think about how determined we are in so many things we do. AI can actually say, with varying degrees of accuracy, how we are going to behave. On certain issues, I do not think that means we do not have freedom. I do think it does tell us something about our default selfishness and self-referentiality.

Brian Green: We are very predictable.

Bishop Tighe: Yes, and if you gamble on that you are more likely to be right than wrong. So how do we promote an idea of human solidarity? Because people talk about the emerging inequality—digitalization may have driven that—and the inequality is not just the enormous wealth of the few against the relative poverty of the many; it also is the access to power of the few against the lack of access to power of the others. Beyond that, there is the question: is there even a sense of shared destiny?

I mean, those to whom wealth is gravitating are interested in using it to promote, maybe, space exploration. I know you are interested in that,[17] and I can see what they are thinking there. But if it is about saving the best of the planet and sending them off to future worlds, rather than, say, the harder thing of having to address human issues here on Earth … the fact that it is easier to deal with the technological challenges is tragic. Who decides that huge resources go to one rather than the other? It is the question of common destiny and, relatedly: do we have that sense of human dignity?

I think there are ways in which AI will teach us to be more alert to the limitations of our freedoms, as it can predict patterns of behavior. But it will also raise huge challenges. For example, if it can tell in advance which men are likely to abuse women. If these men are identified, should we take preventive action? Or might we educate them in advance to help them recover their freedom? I think AI may teach us to be more humble about our understanding of the extent to which we

---

[16] Where Hoffman states his thoughts at the Vatican: Vatican IHD, "Opening Session Part 2: The Common Good: Seeking Shared Values," from "The Common Good in the Digital Age" conference, Sept 26, 2019, uploaded to *YouTube*, Nov 14, 2019, time: 1:01:30-1:03:10, www.youtube.com/watch?v=2FYh_j3OBDg. Hoffman originally stated his idea here: "Reid Hoffman," *The Wall Street Journal*, June 23, 2011, www.wsj.com/articles/SB10001424052702303657404576363452101709880. For influence of this thought, see Robinson Meyer, "The Seven Deadly Social Networks: Every Crime against the Divine Will Has Its Own Corresponding Digital Brand," *The Atlantic*, May 9, 2016, www.theatlantic.com/technology/archive/2016/05/the-seven-deadly-social-networks/480897/. Finally, in 2021 he rethought, clarified, and revised his position to emphasize that impulses towards vice also need to be actively controlled and limited: Reid Hoffman, "Human Nature in Vices and Virtues: An Adam Smith Approach to Building Internet Ecosystems and Communities," *The Knight Foundation*, October 29, 2021, knightfoundation.org/human-nature-in-vices-and-virtues-an-adam-smith-approach-to-building-internet-ecosystems-and-communities/.

[17] Brian Patrick Green, *Space Ethics* (London: Rowman & Littlefield, 2021).

are free. It would be not so much about measuring the freedom that we have, but potentiating that to make good choices.

Brian Green: I like the positive framing you put there. We might appear to lose our freedom, but perhaps we can gain back that and more because we will know the truth.

Bishop Tighe: The whole topic of AI invites us to frame it that way. I mean I can understand why people are saying "'No' to AI making decisions about parole or bail." We know there can be inbuilt biases and that we can get it wrong. We should also know that the judicial system as we have it now may not be as good (effective) as even a semi-good AI system. We have to be careful not to project on humans this extraordinary achievement of all our potentials. We can maybe use AI to help us reflect on who we are and what we are, and understand our patterns of behavior, with the result of us then being able to use that knowledge to grow.

Brian Green: The next big question I was going to ask you is: are there relationships between AI and theology we should be thinking about?

Bishop Tighe: Twenty-five years ago, when some of the stuff on genetic enhancement was coming out, my mentor in the area of moral theology, Maurice Reidy, saw many people in the theological arena simply responding "Oh, you should not play God." Reidy would always retort: "No, you should play God." Our God is a God who created, who created with love and attention. When we begin to deploy these technologies, how can we use them for the good? In other words, we should be as attentive in our stewardship of creation as God was in the act of creation. So that was just to flip that traditional idea that you cannot play God. There are many decisions we have to make.[18] In broad terms, we are at a turning point with all the developments in nanotechnology, biotechnology, information technology, cognitive science, and genetics with AI increasingly driving said developments. These begin to combine together, and ultimately, we are talking about taking human evolution into our own hands. Now, maybe that is overstated, but that seems to be where technology is going, especially in the biological sector. If we begin to do this, questions emerge about the values that should shape it and who should decide.

I think literature can help here. For example, in Kazuo Ishiguro's *Never Let Me Go*, human beings are cloned so that their organs can be given to other people.[19] Ishiguro's question was, in the very beginning:

---

[18] Maurice Reidy, Seminar for Geneticists, Holy Cross College, Clonliffe, 1995. See also Brian Patrick Green, "The Technology of Holiness: A Response to Hava Tirosh-Samuelson," *Theology and Science*, 16, no. 2 (2018): 223–28, where I make a similar point: we should imitate and seek—"play" at—God's holiness (via ethics) and not only "play" at God's power (via technology) or we will, in our unethical power, destroy ourselves.

[19] Kazuo Ishiguro, *Never Let Me Go* (New York: Vintage, 2005).

how did it happen? Because nobody ever wanted to be killing human beings for their organs. Well, what happened was that people wanted to address and cure particular illnesses. A good desire then became inhumane at the close. We need to pause and ask. Rather than let this happen by creeping, well-intentioned incrementalism, we need to initiate a conversation. Who can do that?

For theology, I think the challenge is to determine how we can share our theological insights and translate those into languages that people formed in other disciplines can actually appreciate. At the same time, how do we help people in other disciplines share their insights? I think it will become hugely important in terms of theological formation. We used to insist that people study philosophy before theology. Well now I think people need some awareness of science and technology in preparation for theology, if people are going to be adequately reflecting on our world.

Brian Green: I agree with that a lot. The natural sciences used to be called "natural philosophy," and they would have been part of the philosophy curriculum.

Bishop Tighe: Today, even in terms of images and metaphors, we cannot talk to people if we do not know them.

Brian Green: Exactly. The lack of comprehension becomes a communication problem. Returning to the anthropological side of artificial intelligence, for what else do you think AI might be significant? And [humorously], would you baptize an AI if it asked?

Bishop Tighe: Another important thing is the whole question of ontology. What is the nature of the being of an AI or a robot? As I mentioned previously, we had the neuroscientist Christof Koch at a recent seminar. He was very clear that AI and robots could performatively seem human, but he was very reluctant to ascribe any form of consciousness to artificial intelligences. In other words, you may end up believing you are interacting with a human, but ultimately, the question is: "Is it actually human?" I know you were not being overly serious about the question, but if I had an AI ask me to baptize it, I would not be inclined to.

In functional terms, an AI could participate in a sense of belonging, but I think it might be more akin to a family pet. That does not mean you will not develop feelings about it. I read somewhere that American soldiers expressed grief after robots they used to disarm bombs were damaged by the bombs, and they experienced a sense of loss. I think we need to maintain an ontological perspective rather than just a projection. The ontological issue remains important.

Brian Green: Ontology never goes away.… Moving onto more concrete issues, what teachings of the Church do you think are the most relevant when it comes to thinking about AI?

Bishop Tighe: At the risk of repetition, I think some of the Church's perspective on the incarnational dimension of our lives and

the bodiliness of being human are vital to remember. There is something about the body—"It is not so much that *I have* a body as that *I am* a body"—the old Merleau-Ponty quotation. This is critical. Some Catholic teaching, even in areas of sexuality, also becomes relevant here. In a lot of thinking out there the body is almost reduced to the carrier of the real person. The real person, therefore, is not dependent on the body and could be liberated from the body. I think incarnation and embodiment would be insights we need to bring to counteract the kind of dualism I think can emerge in a lot of thinking on AI. Other areas, more obvious and with immediate applicability, as I said before, are questions about inequality, unemployment, justice issues, and so on.

I think one further issue is something UNESCO highlights: our interactions with robots and AI—which exist almost exclusively to do what we want—could condition how we think about our relationships with real human people.[20] There could develop an expectation that they also exist solely to satisfy my needs. UNESCO was also beginning to look into the issue of the anthropomorphization of AI. I think there are a range of issues about which our Church teaching will have things to say.

Brian Green: Can you say something about what the Church's impact has been on issues related to AI? Have the conversations in which you have been involved turned into action in any ways?

Bishop Tighe: I am not sure anything has turned into direct actions, because I am not sure if the Church in that sense is an actor in the arena, as of now. If we wanted to distinguish a bit about the Church, I think the Church is not simply hierarchies, institutions, and professors of moral theology, it is also individual believers. Catholics, together with other people with different religious backgrounds, or people with no religious backgrounds but with developed ethical thinking, are trying to marry their principles with their work practices. Importantly, we have already seen people working in the AI arena say to the Church: "Help me think through the issues I am addressing in my day-to-day work." And they are beginning to work collaboratively among themselves. I am struck by this as a kind of embodiment of *Gaudium et Spes,* where it says that lay people cannot look to Church leaders for instant answers to every question (no. 43). They cannot expect it, but they can get support in terms of the analysis they can bring to reflect on their responsibilities.

We have a group of Catholic technology leaders in Silicon Valley working on these issues. So we ask: "How do we support and equip people working in the arena to be able to bring their values into

[20] UNESCO, "Draft Text of the Recommendation on the Ethics of Artificial Intelligence," UNESCO website, November 22, 2021, esp. § 128 and 129, unesdoc.unesco.org/ark:/48223/pf0000379920.page=14.

conversation?" Not in a sectarian way, but to contribute to the overall purpose of the company, and try to ensure that AI will be in service of humanity: what is true, what is good, and so on. Because they are the ones who are there. In terms of practical things, in the discussions with secular Silicon Valley leaders, what has emerged is mutual respect and appreciation that people working in this field, at the development phase, have good intentions. They also have commercial and other intents, but they have fundamentally good intentions and they certainly are anxious not to do any harm.

From the other side (and they were the ones who did the inviting initially), I would say there has been a growing awareness that our tradition has insights into what it means to be human and what it means to live in society, relevant to them and their concerns. This dialogical context, as the relationship gets better and more mature, gradually allows for a more frank and open critique of one another's positions. A more robust discussion develops, and that is an achievement.

Brian Green: I do want to push a little bit more just, because I think most people have no idea these conversations are even going on in the first place. So, I wonder if you could give more specific details about some discussions in which the Vatican has been involved. Or if you could mention ways in which the conversation might have progressed because of the Vatican perspective, or things which might be unique to the Vatican's contribution, even if it is more conversational than active.

Bishop Tighe: The initial conversations with secular leaders tended to be very much centered around shared texts, which were essentially articles chosen by people from Silicon Valley, which they felt even people with no technical background would be able to read and gain a sufficient understanding of the issues. That educational purpose was there to begin with. A lot of discussions then were focusing in and around texts. To some extent that has remained the model, but the sophistication of texts has improved.

Equally, on the Vatican side, where there are quite a few academics, we found some very helpful ideas coming from Scholasticism and Thomas Aquinas. Some of us had to translate those ideas into terms more intelligible for the people on the other side! There was a lot of bridging, and learning each other's languages, and trying to understand, that has been enabled by the experience of working together, by the social dimension. A lot of really good conversations happen not so much at the table, where we have our fixed points of discussion and articles we are trying to follow and debate, but in the margins of those, where people raise questions and begin to express and explore ideas that maybe they would not take to the full table.

The fruits of that would be found in a conference like the one held in 2019, in being able to bring to the table at a Vatican conference people like Reid Hoffman from LinkedIn, Mitchell Baker from

Mozilla, and others. They were able to come and be part of a public engagement in that area. In the same way, there are other people, particularly Eric Salobir, who have been brought into much more corporate environments. Of course, there are more; this is not the only show in town.

As these conversations progressed, I also realized that we needed to tap into far wider Church sources, and that is where the Pontifical Council for Culture's Center for Digital Culture came from: an intuition that real knowledge and expertise and insight could be brought from a broader Church perspective. We decided to engage with people coming from the global network of Catholic universities, looking for people with the competencies to think about these topics. They are also a resource that will eventually help the Vatican become more sophisticated in how it thinks about these issues.

As you know, we have been working in these academic groups now for about two years and have never been able to meet in person because of COVID-19. We started meeting online instead, and it is beginning to help us tap into a wider range of people, not just in the United States and Europe, but also in Asia, Latin America, and beyond. It is also showing us that there is a lot of thinking and reflection available to us that may not be coming from an explicitly Catholic context but out of similar value systems, which can aid communication.

Brian Green: Moving towards the future: what do you see as the future of the Church's engagement with AI or the future of the PCC's work on AI?

Bishop Tighe: I think the near future is about leveraging the interest there is in the Church on AI and ethics in non-Church environments, to facilitate closer thinking and reflection. In the long term it is also about developing, on behalf of the Church—this could involve the Council for Culture and the Center for Digital Culture working with the Congregation for Catholic Education, which has oversight of Catholic universities—a formal invitation to Catholic universities to ethically reflect on AI and technology in general. It would really help if we could bring this more to the forefront of what they are doing—this is a difficult thing—because we need to develop people sufficiently fluent in the technological area and sufficiently in tune with its culture who can then credibly bring insights coming from positions of faith into those discussions.

I mean the danger is what Christof Koch said to us, which is that people will still be talking about this, and one day it will have happened. We will get left behind; the talk will have had no effect. This problem is one of the reasons why it is important that the Holy See bring the insights of our tradition to the international organizations, such as the UN and UNESCO, where it has representation. We cannot but acknowledge that this is a challenging moment for generating

international co-operation, judging by what we saw at COP26 in Glasgow in the environmental conversations. The global community is in general very fractured, and AI is another issue upon which it is likely to fracture.

I do think there is an appreciation of what the European Union is proposing to do, which is, rather than coming up with a comprehensive regulation—the aim of their initial effort—to look instead at the ethics of AI and come up with a much more prudential line-drawing exercise relating risk to the degree of regulation. So, the more risky the activity, the more it has to be regulated and controlled; this rather simple insight does help to set certain standards.

Brian Green: I only have two questions left. The first is, and you were just touching on this: what are your hopes and fears for AI and the development of AI going forward?

Bishop Tighe: My hope is that the undoubted potential of AI to process large data sets and deal with issues of complexity would be a very helpful tool we use to model different options, particularly on issues around the environment and the like. I am not saying technology is the only answer, because it is not. Technology, data, information, and proper understanding of the realities in which we are, I think, are very important. I would also like to think that AI would allow for a way of aligning humanly valuable applications with the more commercially viable ones, so there would be an alignment between the incentives for companies and the nobility of what they are trying to pursue in their activities.

When we look at digitalization and the internet in general, I think it is fair to say that a lot of the most monetized things have not necessarily brought forth the most noble aspects of its potential. Will AI in the long term be used as something to help us address real global problems, or will it be something used to satisfy rather immediate needs of a privileged minority, likely to be paying for that? My hope is to see what the positive is for AI, and then the negative is the failure to realize the positive, rather than getting down the track of the particular dangers.

Brian Green: Once again, you are leading with the positive. My last question is: do you have any final thoughts or anything else you would like to add?

Bishop Tighe: I would like to say, at a personal level, that for the field of moral theology and people working in professional ethics, there is enormous potential to make a real contribution to this conversation. Avoid the temptation of being the external experts who offer extrinsic solutions, become the people who facilitate the actual decision makers in thinking ethically. As I have said, I am not a highly speculative thinker, but I think many technologists are even less speculative. They want to deal with what is tangible, real, and can be measured, and yet here we have to go into questions not so amenable to that

approach. The questions are messier, and they require negotiation and discussion and engagement with different positions.

I have mentioned Kazuo Ishiguro before. He has a recent novel on AI called *Klara and the Sun*.[21] I was struck personally by an interview with Ishiguro in which he talked about growing up with his father, who was a scientist working on climate issues, long before it was fashionable.[22] Ishiguro was always very impressed by how the scientific community reasoned, and how people manage to formulate hypotheses, which stood as long as they were validated, and fell once they were disproven, and yet the community worked as one, together, in that. To have advanced an ultimately wrong hypothesis might have been helpful and there was nothing personal about it. I think we have to have a similar way of thinking about human issues. Ishiguro talks about what he calls "proper truths": really important truths.[23] These enable people to think and reflect together and discern what is really going to support humanity.

This brings me back to the idea that ethics is a method and approach to our dilemmas in life. In science and technology, rather than offering extrinsic solutions, it helps to see ethics as intrinsic to what they are doing and enabling them to be more comfortable in addressing the not-so-black-and-white questions, the not-so-binary issues. I do think sometimes the default position for many scientists is to end up working with a consequentialist moral theory because it kind of seems scientific. One of the problems, then, if that is your approach, is that you displace what does not fit into the theory. What cannot be measured gets excluded. The system of measurement is what we need to question. We have to ask: how do we measure what is humanly good, what is globally attractive, and how do we do that in an inclusive way? There is a lot more to be said.

Brian Green: There is much more to say and that is a wonderful place to conclude. This has been a fantastic interview. I really appreciate it, and all the time you have taken to promote work on this subject.

Bishop Tighe: Thank you. M

Bishop Paul Tighe is the Secretary of the Pontifical Council for Culture. He was born and raised in Ireland and graduated from University College, Dublin, with a degree in Civil Law. He was ordained a priest in 1983 after studying at Holy Cross College, Dublin, and at the Pontifical Irish College in Rome. He then studied moral theology at the Pontifical Gregorian University.

---

[21] Kazuo Ishiguro, *Klara and the Sun* (New York: Knopf, 2021).

[22] Steve Paikin, "Kazuo Ishiguro: A Nobel Novelist Searches for Hope," *The Agenda*, on *YouTube*, Mar 10, 2021, min. 19–25, www.youtube.com/watch?v=5DmZqJW8nWw.

[23] Steve Paikin, "Kazuo Ishiguro: A Nobel Novelist Searches for Hope."

In 1990 he became lecturer in moral theology at the Mater Dei Institute of Education in Dublin and at Holy Cross College and was appointed head of the theology department in 2000. Bishop Tighe was named Director of the Communications Office of Dublin Diocese in 2004, and he established the Office for Public Affairs in 2005, with the purpose of engaging the Diocese with public institutions and civic society. Pope Benedict XVI appointed Tighe as Secretary of the Pontifical Council for Social Communications in 2007 and Pope Francis appointed him titular Bishop of Drivastrum and Adjunct Secretary of the Pontifical Council for Culture in 2015. At the Council, where he now serves as Secretary General, he follows questions related to digital culture (the impact of technology on social and political discourse), ethics, and contemporary literature. Brian Patrick Green is Director of Technology Ethics at the Markkula Center for Applied Ethics at Santa Clara University.

Journal of Moral Theology, Vol. 11, Special Issue 1 (2022): 232–249

Epilogue on AI and Moral Theology: Weaving Threads and Entangling Them Further[1]

Brian Patrick Green

ARTIFICIAL INTELLIGENCE TECHNOLOGY WILL affect religion and moral theology, and these impacts will run the full spectrum from negative to positive, with neutral, mixed, and ambiguous changes thrown in. Artificial intelligence is human intelligence to the next degree, perhaps eventually a seemingly infinite degree, Google CEO Sundar Pichai declaring it "more profound" than fire, electricity, or the internet.[2]

AI already has had effects on popular "moral theology," with various internet groups declaring that AI will become god-like,[3] or that we are living in a computer simulation,[4] and from these deriving behavioral guidance, to the point of worshipping AI[5] or fearing Roko's Basilisk: the wrath of the coming AI god who will torture you for not

[1] Parts of this paper are based on Brian Patrick Green, "Some Ethical and Theological Reflections on Artificial Intelligence," presented at the Pacific Coast Theological Society conference, Graduate Theological Union, Berkeley, CA, November 3, 2017, www.pcts.org/meetings/2017/PCTS2017Nov-Green-ReflectionsAI.pdf. The ethical content of that paper was published as "Ethical Reflections on Artificial Intelligence," *Scientia et Fides* 6, no. 2 (2018): 9–31, dadun.unav.edu/bitstream/10171/58244/1/01.pdf; the more theological content is provided here.

[2] Amol Rajan, "Google Boss Sundar Pichai Warns of Threats to Internet Freedom," *BBC News*, July 12, 2021, www.bbc.com/news/technology-57763382.

[3] See Way of the Future Church website, "What is this all about?," November 16, 2017, web.archive.org/web/20171116133733/http://wayofthefuture.church/; for an interview with Anthony Levandowski about his church, see Mark Harris, "Inside the First Church of Artificial Intelligence," *Wired*, November 15, 2017, www.wired.com/story/anthony-levandowski-artificial-intelligence-religion/; for an article on the church closing, see Kirsten Korosec, "Anthony Levandowski Closes His Church of AI," *TechCrunch*, February 18, 2021, techcrunch.com/2021/02/18/anthony-levandowski-closes-his-church-of-ai. The church lives on through fans on Twitter: Way of the Future (AI Church)@wayofthefuture_, "A Sufficiently Advanced Artificial Intelligence Would Be Indistinguishable from God," (W.O.T.F. fan account), twitter.com/wayofthefuture_?lang=en.

[4] Nick Bostrom, "Are You Living in a Computer Simulation?," *Philosophical Quarterly* 53, No. 211, (2003): 243–55, www.simulation-argument.com/simulation.html.

[5] See Mark Harris, "Inside the First Church of Artificial Intelligence," *Wired*, November 15, 2017, www.wired.com/story/anthony-levandowski-artificial-intelligence-religion/.

having done enough to bring about its advent.[6] It is only a matter of time before these behavioral impacts become more widespread and intense.

This "moral theology" of artificial intelligence will require the attention of Christian moral theologians, if for no other reason than by not doing so we will be out of touch with contemporary culture. There are significant insights to be gained from thinking about moral theology and AI, and teasing apart the similarities and dissimilarities will be beneficial to our thinking as moral theologians living in the contemporary world.

Christian theologians and ethicists have a part to play in this secular conversation on AI. We might feel left out and certainly have a lot to do in order to catch up with the secular conversation on both AI and AI ethics, but we also have fundamental insights to share—insights that are wanted and needed to provide guidance on how to use AI and protect human dignity. Awareness of our limitations should encourage us to humbly want to learn and do more.

In this epilogue I will bring together threads of ideas from this special issue and add some additional yarn as well, highlighting the contributions of the authors and how much more there is to say. I will not claim to be able to untangle all knots, limiting myself to categorizing the material into anthropological, theological, and ethical threads.

More Anthropological Threads

As many of the papers in this volume note, AI raises fundamental questions about humanity. As we externalize human intelligence and develop it beyond our own comprehension (as has already been done[7]) we might well ask if we are making ourselves obsolete or inferior to the works of our hands. Questions of consciousness, mind uploading, idolatry, and the role of humanity in God's creation come to mind, among others.

Can Machines Be Conscious?

Artificial consciousness may seem like a very theoretical and impractical concern, but it has immense relevance for moral theology for at least three reasons. First, conscious machines might reasonably seem to count as moral persons, and then would need to be treated as such. Second, conscious machines would need to behave ethically and could be judged for their ethical behavior. Third, conscious machines

[6] A number of people have taken this idea very seriously, see "Roko's Basilisk," www.lesswrong.com/tag/rokos-basilisk.

[7] See for instance Bob Yirka, "Computer Generated Math Proof is Too Large for Humans to Check," *Phys.org*, February 19, 2014, phys.org/news/2014-02-math-proof-large-humans.html.

could perhaps be in need of salvation, which would raise some questions for Christianity, to say the least.

The papers in this special issue engage all three of these concerns. Mark Graves's article directly engages the first two reasons in significant depth. He realizes the importance of the question he is tackling: "Words such as 'moral,' 'conceptualize,' 'actor,' 'reckon,' etc., we typically reserve for the behaviors of self-conscious agents like humans are, and while I do not rely on that interpretation here, I leave open the possibility that AI might someday attain that status."[8] While the possibility of AI becoming self-conscious is left open, and his article further investigates how this might become more likely in the context of moral self-reckoning, Graves's is a lonely voice in this volume. Graves's connection of action to consciousness is a clear one, however; after all, if something is going to interact with the world in an ethical way, it must be able to delineate itself, the world, how it affects the world through the actions it takes, how the world might react in return, and the relevance of this for yet further impacts. While "reckoning" these things might be possible without self-consciousness, some form of "awareness" would be necessary for this sort of machine activity, even if a completely non-conscious one, alien to humankind.

In the conversation essay, however, we see several participants, such as Marga Vega and Anselm Ramelow, OP, clearly state that they see AI consciousness as highly improbable, if not logically impossible.[9] On ethical grounds, Levi Checketts is highly skeptical of associating AI and consciousness—at least as many currently consider it in society—since it not only falsely elevates the machine, but also falsely degrades humanity.[10] Building upon Emmanuel Levinas, Roberto Dell'Oro argues against rationalist and empiricist understandings of personhood, leaving no possibility for AI consciousness.[11] Jordan Joseph Wales argues similarly: we are the ones seeing the world, AI merely sits between us and the world, helping us to see; it does not itself see in a conscious sense.[12] In his interview, Bishop Paul Tighe

[8] Mark Graves, "Theological Foundations for Moral Artificial Intelligence," *Journal of Moral Theology* 11, special issue 1 (2022): 182–211.

[9] Brian Patrick Green (ed.), Matthew J. Gaudet (ed.), Levi Checketts, Brian Cutter, Noreen Herzfeld, Cory Labrecque, Anselm Ramelow, OP, Paul Scherz, Marga Vega, Andrea Vicini, SJ, and Jordan Joseph Wales, "Artificial Intelligence and Moral Theology: A Conversation," *Journal of Moral Theology* 11, special issue 1 (2022): 13–40.

[10] Levi Checketts, "Artificial Intelligence and the Marginalization of the Poor," *Journal of Moral Theology* 11, special issue 1 (2022): 87–111.

[11] Roberto Dell'Oro, "Can a Robot Be a Person? De-*Facing* Personhood and Finding It Again with Levinas," *Journal of Moral Theology* 11, special issue 1 (2022): 132–56.

[12] Jordan Joseph Wales, "Metaphysics, Meaning, and Morality: A Theological Reflection on AI," *Journal of Moral Theology* 11, special issue 1 (2022): 157–81.

finds the idea of an AI legitimately being able to request baptism (as in being free, rational, conscious, and willing, and not merely simulating those traits) to be doubtful.[13] A majority of voices, then—at least in our small group—seem to be against the idea of the possibility of AI consciousness.

One problem remains. As Brian Cutter notes, how could we ever really know if an AI were conscious and, beyond that, determine its moral status?[14] Consciousness is, at its very ground, inscrutable. We believe other humans have experience, because we ourselves do. We grant that belief in other minds based on the very reasonable assumption that humans are in many ways similar to one another in our experience of the world. But we are not computers. We cannot as easily share that assumption with them. How could we know if a computer experiences consciousness, since we are not computers ourselves?

Or are we computers ourselves? As Checketts notes, this is the next problem: by thinking that machines can become conscious we could simultaneously reduce our consciousness to the level of mechanism. The moral theological implications of this perspective are not only dehumanizing, which poses a clear ethical threat, they also make religion vulnerable to a form of replacement: the belief in mind uploading and technological afterlife. Notice I do not say the vulnerability is in the reality of *mind uploading*—the danger is in the mere *belief* in mind uploading, because that belief is enough to provoke certain human thoughts and behaviors. The practical plausibility of mind uploading (quite low) actually has little relevance, at least at this point. Faith in technological and moral progress is enough to motivate this religion.

Mind Uploading: Will Technology Leave Our Brains and Religion Behind?

In the conversation paper, Checketts, Noreen Herzfeld, and Cory Labrecque note that the idea that human minds could be detached from human bodies and placed into silicon ones is not only metaphysically questionable, but also theologically problematic.[15] Bishop Tighe also notes the "speculative" nature of uploading, preferring to stay closer to reality.[16] But because human minds run on stories and beliefs, the latter are enough to impact society. For human minds, belief is reality. And some technological advances are worth tracking, Neuralink's

[13] Brian Patrick Green, "The Vatican and Artificial Intelligence: An Interview with Bishop Paul Tighe," *Journal of Moral Theology* 11, special issue 1 (2022): 212–31.
[14] Green, Gaudet, Checketts, Cutter, Herzfeld, Labrecque, Ramelow, Scherz, Vega, Vicini, and Wales, "Artificial Intelligence and Moral Theology," 13–40.
[15] Green, Gaudet, Checketts, Cutter, Herzfeld, Labrecque, Ramelow, Scherz, Vega, Vicini, and Wales, "Artificial Intelligence and Moral Theology."
[16] Green, "The Vatican and Artificial Intelligence," 212–31.

new brain-computer interface being just one.[17] We are already well-over a decade into the era of humans and animals moving cursors on a screen with their minds.[18]

Brain-computer interfaces demonstrate that human thought involves our brains. This is not an issue for most versions of Catholic theology, such as Thomistic hylomorphism, which argues for a fundamental connection between matter and form, but it does present interesting challenges if large sections of the brain could be replaced with external computing and information storage.

These sorts of developments will challenge some of the supporting cultural assumptions of particular religions and theologies, for example, the nature of immortality, prayer, and the reality of heaven. If, for example, very realistic simulations of people could be created, including family members or historical figures such as the saints, what would this mean in the context of immortality, prayer, and heaven? Would the "*aether*" in which these AIs "lived" and had their being become like heaven, where the deceased go to carry on a simulated existence? Would our texted, verbal, or virtual reality inquiries of them become our prayers for intercession?

Even short of virtual immortals in a virtual heaven, such devices as neural prostheses, brain-computer interfaces, and so on, throw into question some of our deepest assumptions about reality and religion, not to mention anthropology, ethics, and politics.[19] Humans seem to have an innate body-soul "folk-dualism," which of course has crept into Christianity as the idea of heaven filled with disembodied souls playing harps on clouds.[20] The biblical resurrection of the dead is, of course, a very different proposition from this folk conception. Theology might actually be better placed to take on this more materialist reality than we realize; the folk-dualists should really have trouble with it. Unless, of course, the dualism becomes one of hardware and software: a metaphor that has been in use for years.

How many assumptions of Christian or theistic faith will be rendered confused or unintelligible to contemporary culture? In part due to a technologically divergent cultural context, I am seeing some students having great difficulty understanding basic theistic ideas. The

---

[17] "Monkey MindPong," *Neuralink* website, n.d. (uploaded to *YouTube* April 8, 2021), neuralink.com/blog/monkey-mindpong/.

[18] Gopal Santhanam, Stephen I. Ryu, Byron M. Yu, Afsheen Afshar and Krishna V. Shenoy, "A High-Performance Brain–Computer Interface," *Nature* 442 (13 July 2006): 195–98, www.nature.com/articles/nature04968.

[19] Charles E. Binkley, Michael S. Politz, and Brian P. Green, "Who, If Not the FDA, Should Regulate Implantable Brain-Computer Interface Devices?," *AMA Journal of Ethics* 23, no. 9 (September 2021): E745–49.

[20] Edward Slingerland and Maciej Chudek, "The Prevalence of Mind–Body Dualism in Early China," *Cognitive Science* 35 (09 June 2011): 997–1007, onlinelibrary.wiley.com/doi/full/10.1111/j.1551-6709.2011.01186.x.

difficulty can be captured in the real student question: "Does heaven have free Wi-Fi?" (in a later retelling, a Dominican friar quipped to me in return: "In heaven *you are* Wi-Fi"). Additionally, besides "confused" and "unintelligible," there is a third option: hijacked. That is what technology is doing when it creates substitutes for prayer, immortality, and heaven: hijacking Christian ideas and materializing them into machines.[21]

So, is our religion becoming outdated? Can Western religion be updated or has it run its course? Are we just raising a feral generation quite capable of reading text on a screen or performing great feats at video games, yet unable to understand even the basics of human life, relationships, and well-being, let alone history, philosophy, or culture? This delight in the work of our own hands should remind us of an old sin: idolatry.

Humans as Creators of a God or Idolatry—"God is Like an AI" (in a Bad Way)

Ramelow and Herzfeld both warn about the idolatry of technology[22]; the problem already is, I believe, much bigger than we might expect. For example, Anthony Levandowski's desire to create an AI to function as a god manifested as his Way of the Future Church.[23] This intoxication with power and idolatry of technology will not turn out well, and almost certainly lead to disaster.[24] Herzfeld notes that the creativity of the image of God in us can all too easily become distorted and go astray.[25] We should not idolize technology, just as we should not idolize money, power, or other things. But being humans with a predisposition towards sin, that is in fact what we do. AI will just be the next big thing.

As will be explored more below, the "God as AI" metaphor might be helpful to our understanding of God (though with the limitations of any analogy), but the reverse, "AI as God," should frighten us immensely. We cannot make God.[26] Any "God" we could make would be a terribly inferior "god" indeed. The expectation that we could

---

[21] The idea of our religion being hijacked should give us some solace, for it means something is worth hijacking. Our job should be to determine exactly what that is and present it to the culture in its authentic form.

[22] Green, Gaudet, Checketts, Cutter, Herzfeld, Labrecque, Ramelow, Scherz, Vega, Vicini, and Wales, "Artificial Intelligence and Moral Theology," 13–40.

[23] Olivia Solon, "*Deus ex machina*: Former Google Engineer Is Developing an AI God," *The Guardian*, September 28, 2017, www.theguardian.com/technology/2017/sep/28/artificial-intelligence-god-anthony-levandowski.

[24] Brian Patrick Green, "The Technology of Holiness: A Response to Hava Tirosh-Samuelson," *Theology and Science* 16, no. 2 (2018): 223–28.

[25] Green, Gaudet, Checketts, Cutter, Herzfeld, Labrecque, Ramelow, Scherz, Vega, Vicini, and Wales, "Artificial Intelligence and Moral Theology," 13–40.

[26] Brian Patrick Green, "Transhumanism and Roman Catholicism: Imagined and Real Tensions," *Theology and Science* 13, no. 2 (2015): 187–201.

somehow create a god might reflect something of our feeling of entitlement and ingratitude at the situation in which we find ourselves. When the inevitable bubble of hubris bursts, we may find among those left alive a newfound appreciation for the real God. In times of trial and failure we turn to God, and in the absence of trial and failure perhaps we may tend not to. Jewish and Christian ethics both emphasize humility as inoculation against hubris. If one does not try to illegitimately raise oneself up to Babel-like heights, one cannot fall from those heights. We are called to humility, but not humiliation.[27]

There is a danger in mythologizing or theologizing technology. Religious language is a constant part of discussions about AI, for the reasons noted here, and more. Despite these comparisons, we must make absolutely sure we do not come to see our metaphors and thought-devices as reality. Humans are tool users and tool makers, but we should not become tool worshippers. Our capacity to see teleology in tools and teleology in our lives and God may take root in the same cognitive abilities,[28] but they should not be confused. God is not a tool and tools are not gods. The mythologization of technology leads us away from reality.[29]

In expressing our desire to create we express a God-given talent. God created humankind, and now we create a world full of tools, including AI tools. Do our multifarious creations reflect well on us? Do we as creations reflect well on God?

In our imaginations we perceive AI to be both our slave and our God. Both of these mythologizations are terribly misguiding. AI cannot be our God or even a god. Considering AI a slave just perpetuates the mindset of a slave master in our own minds, habituating us towards vice. Let us purge ourselves of both these impulses and see AI as what it is: complex math which can aid human intelligence.

God's Creativity Returns: Humans as God's "AIs"

There is a long tradition of analogizing God's creation and human creation. With God understood as artificer, nature becomes God's artifact and, likewise, our artifacts become analogues of natural objects.[30] As we create artificially intelligent systems, we can perhaps

---

[27] Hans Jonas, *The Imperative of Responsibility* (Chicago: University of Chicago Press, 1984), 202.

[28] Brian Patrick Green, "Teleology and Theology: The Cognitive Science of Teleology and the Aristotelian Virtues of *Techne* and Wisdom," *Theology and Science* 10, no. 3 (13 August 2012): 291–311, www.tandfonline.com/doi/abs/10.1080/14746700.2012.695247.

[29] Kevin Kelly, "The AI Cargo Cult: The Myth of a Superhuman AI," *Wired*, April 25, 2017, www.wired.com/2017/04/the-myth-of-a-superhuman-ai/.

[30] See for example Wisdom 7:22 and 8:5–6, Matthew 13:55, Mark 6:3, and Hebrews 11:10. Further, Simon Francis Gaine, OP, points out that both Augustine and Thomas Aquinas fruitfully build upon this tradition, see Thomas Aquinas, *Summa Theologiae*,

place ourselves in an analogous position to God—for good and evil. For evil if we try to compete with and rival God.[31] For good if we try to cooperate with God, and play our role as images of God and stewards of God's Creation.

One thing humans do with AI systems is delegate authority. An algorithm absorbs data and learns to identify images, words, and even the beginnings of "concepts."[32] In cyberdefense, an AI model "watches" for cyberattacks and automatically responds. In these cases, we humans delegate this authority because we ourselves are not as able to do the work. We can absorb images, words, and concepts better than AI, but a machine learning model is much more effective than inputting all the data into a computer model by hand, which we cannot do with very large data sets. Relatedly, when it comes to cyberdefense, humans are simply not fast enough.

Unlike AI, God did not create us to complete tasks delegated to us because we are somehow "better" at these than God. God is self-limiting and in this self-limitation God gives us true freedom and the ability to truly love that comes along with that freedom: something that we cannot yet give, and likely (in my opinion) never will be able to give, to AI. The authority God delegated to us was the authority to freely love, not because God *cannot* do that, but because God *can*, and one reasonable effect of that love is to be fruitful and give this ability to love to others. This is the image of God in us and the role, then, of technology is to empower us to love more completely, as Bishop Tighe,[33] Andrea Vicini, SJ,[34] and Wales all note.[35]

I, q. 14, a. 8; q. 27, a. 1, ad. 3; q. 39, a. 8; q. 44, a. 3; q. 45, a. 6; III, q. 3, a. 8; Thomas Aquinas, *Summa Contra Gentiles*, IV, 13; IV, 42; Augustine, *De Trinitate* 6.10.12; Paul T. Durbin, "Aquinas, Art as an Intellectual Virtue, and Technology," *New Scholasticism* 55 (1981): 265–80; Simon Francis Gaine, "God Is an Artificer: A Response to Edward Feser," *Nova et Vetera* 14 (2016): 495–501; Francis J. Kovach, "Divine Art in St. Thomas Aquinas," in *Arts libéraux et philosophie au Moyen-Âge* (Paris: Vrin, 1969), 663–71. These texts are cited in Brian Patrick Green, "The Catholic Church and Technological Progress: Past, Present, and Future," *Religions* 8, no. 6 (2017): 106.

[31] René Girard, "Mimesis and Violence," *The Girard Reader*, ed. James G. Williams (New York: Crossroad, 1996), 9–19; citing René Girard, "Mimesis and Violence: Perspectives in Cultural Criticism," *Berkshire Review* 14 (1979): 9–19; cited in Brian Patrick Green, "A Catholic Perspective: Technological Progress, Yes; Transhumanism, No," in Arvin M. Gouw, Brian Patrick Green, and Ted Peters, eds., *Religious Transhumanism and Its Critics* (Lanham, MD: Lexington, 2022), 146, 152.

[32] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah, "Multimodal Neurons in Artificial Neural Networks," *Distill*, March 4, 2021, distill.pub/2021/multimodal-neurons/.

[33] Green, "The Vatican and Artificial Intelligence," 212–231.

[34] Green, Gaudet, Checketts, Cutter, Herzfeld, Labrecque, Ramelow, Scherz, Vega, Vicini, and Wales, "Artificial Intelligence and Moral Theology," 13–40.

[35] Wales, "Metaphysics, Meaning, and Morality," 157–81.

In our creative processes, then, we can cooperate with God. If God created us to love, by delegating the capacity for love to us, we should with our technology express God's love onwards and into the rest of creation. That should include use of AI. If we cooperate with God, we will create a better world. If our creations—our AI—cooperate with that Divine project and do not compete with it, likewise things should get better.

MORE THEOLOGICAL THREADS

AI also raises theological questions. Here are just two: analogizing God and AI for the sake of theological reinterpretation, and AI-enhanced theological reflection.

God as an AI or Model Architect: "God is Like an AI" (in a Good Way)

Treating AI as god is clearly idolatrous. But analogizing God—the ineffable ground of all being—to AI, while wholly insufficient, might yield some interesting insights for us mere mortals.

The idea of a superintelligence guiding us and helping us is not just a technological dream—it is a theistic axiom. Bishop Robert Barron once noted that the Waze app guided him through Los Angeles in a way that he found to make utterly no sense—until he arrived at his destination and another person explained that Waze had routed him around a major traffic jam.[36] Waze had access to more knowledge and understanding of the situation than Barron, who just had to take it on faith.

As we attempt to create our own superintelligent tools, our experiences with them will potentially teach us something about God. For example, Nick Bostrom has proposed the simulation hypothesis, where humans live in a computer simulation.[37] Bostrom's idea was quickly turned into the New God Argument by Mormon transhumanist Lincoln Cannon, thus demonstrating the potential fruitfulness of the conversation between technology and religion.[38] While this New God Argument might not do much for non-Mormon theology, an entire Mormon Transhumanist Association has sprung up on its basis.[39] The Mormon path allows for the idea of a superintelligence that humans can create which, in turn, opens up the idea of a superintelligence that created humans. This gives new metaphors for understanding God and increases the plausibility of God's existence, at least according to

[36] Bishop Robert Barron, "The 'Waze' of Providence," *Word on Fire*, December 1, 2015, www.wordonfire.org/articles/barron/the-waze-of-providence/.
[37] Nick Bostrom, "Are You Living in a Computer Simulation?," *Philosophical Quarterly* 53, no. 211 (2003): 243–55, www.simulation-argument.com/simulation.html.
[38] Lincoln Cannon, "The New God Argument," n.d., new-god-argument.com/.
[39] See Mormon Transhumanist Association website, transfigurism.org/.

some people. Even the aggressive atheist Sam Harris has had to admit that the simulation argument increases the plausibility of religion in ways he did not expect.[40]

As Christian theologians we cannot quite afford the Mormon transhumanist idea of "superintelligences all the way down," but we can find other fruitful connections. For some people this superintelligence might take the form of a deistic "divine watchmaker" who created a universe and then left it to run down. For others it might instead increase the plausibility of theism, clarifying the idea that "God works in mysterious ways" because God, like AI, is much smarter than we are. We should not underestimate the ability of a powerful metaphor to capture the human mind. I predict that the "God is/as AI" metaphor will become a powerful one. We do need to be aware, however, that God is not an "artificial" intelligence, but rather a Divine one, so perhaps the shorthand for God ought to be DI for the one Divine Intelligence.

AI-Enhanced Theological Reflection—Can AI Help Us Know God?

Just as AI will have a practical effect on research and education, so too will this include theology. What will AI teach us about God? If we feed an AI everything to know about God will it tell us that God exists with X probability, that God does not exist with Y probability, or that the question remains inconclusive? What other (perhaps more conclusive) questions might we ask of a theologically-trained AI?

AI gives us the opportunity to comprehensively analyze more data than any human could ever understand. Just as humans are biased, so too are the artifacts we make.[41] If an AI—perhaps surprisingly—concludes that God is likely to be real, will its creators then re-train the program to come to a different conclusion? If it concludes that God does not exist will the creators then re-train it to agree that God exists? These questions are already posed to AI, and have been for years, through searches on Google, Yahoo, Bing, and Wolfram Alpha (which when asked, "Does God exist?," states: "I'm sorry, but a poor computational knowledge engine, no matter how powerful, is not capable of providing a simple answer to that question"[42]).

Simpler matters as examining scholarly ideas and writing scholarly papers could be assisted by AI. Has another scholar misinterpreted your favorite theologian? Data mine the theologian's works for the

---

[40] Sam Harris, "Should We Be Mormons in the Matrix?," *Sam Harris's Blog*, www.samharris.org/blog/item/is-religion-true-in-the-matrix.

[41] See Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," *ProPublica*, May 23, 2016, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; BBC Editors, "Google Apologizes for Photos App's Racist Blunder," *BBC News*, July 1, 2015, www.bbc.com/news/technology-33347866.

[42] See Wolfram Alpha website, www.wolframalpha.com/input/?i=does+god+exist.

ideal text to refute them. Do you think a text would be more correctly translated if it were modernized? Run the text through contemporary translation software. Wondering what an ancient theologian might say in response to current ethical issues? Perhaps an AI simulacrum of that person could extrapolate from their previous writings.

Long ago Blessed Raymond Llull, in 13th century Majorca, dreamed of a computational machine, the *Ars Generalis Ultima*, which could answer any question about theology. He spent decades prototyping, refining, and demonstrating it. While many of his contemporaries were unimpressed, he is now recognized as a significant figure in the history of combinatorial logic and computational theory.[43] With AI, we finally approach capabilities that could make Llull's dream become a reality. Will we choose to try to do it? How do our great technologies and low ambitions compare to those of one Majorcan man 700 years ago?[44]

The transhumanist movement captures the attention of more and more young people in the Western world. As Christianity declines, a religion of technology is rising. While Jesus Christ, God, and heaven now seem abstract and distant, technology is before us and constantly growing, with seemingly no end to its ambitions. Perhaps if Christianity showed an equivalent level of vision and ambition—with a heroic moral focus using the gifts of technology when appropriate—it might rightfully regain the attention lost. To paraphrase the 1970s television show *The Six Million Dollar Man,* "we can rebuild it, we have the technology."[45] More than that, we have the morals and the mandate: to love God and neighbor. Good is to be done and we are the ones to do it. Technology can aid but not replace us in this work.

MORE ETHICAL THREADS

There are as many ethical issues related to artificial intelligence as there are ethical issues related to human intelligence. There is no end to ethical threads. However, there are some particularly important, and a few less-discussed ones—we might want to mention: AI use and ethics, AI shifting us from participants to observers of the world, and the imbalance of power and ethics.

---

[43] Raymond Lull, *The Ultimate General Art*, Labirinto Ermetico (English) website, www.labirintoermetico.com/12arscombinatoria/Llull_R_Ars_Generalis_Ultima_(tr._inglese).pdf.

[44] Brian Patrick Green, "What Has Technology to Do with Theology? Towards a Theology of Technology," presented at the "What Has Athens to Do with Jerusalem?" conference, Dominican Colloquia in Berkeley, CA, July 16–20, 2014.

[45] Kenneth Johnson (producer), "Opening Sequence," *The Six Million Dollar Man*, (Universal City, CA: MCA TV / NBC Universal, 1973–1978), on *YouTube*, www.youtube.com/watch?v=BthNjd_jUl4.

Can AI Be Used Ethically? Can AI Be Ethical? Can Humans Be Ethical?

The papers in this volume debate whether AI can be used ethically or even "be" ethical. Graves gives a clear plan for how to make artificial systems act in a way that could be consistent with human morality—in this case it might even *be* ethical, not just *be used* ethically.[46] Vega also mentions the need to build in ethics early and not only concentrate on use.[47]

Most authors tend to avoid "being" language for AI and ethics, focusing instead on its use. Bishop Tighe argues for a positive vision for AI used properly so that we only later perhaps say "no" to bad uses which fall short of the positive vision.[48] Herzfeld gives a strong argument against creating autonomous weapons; any such use would directly threaten many cherished moral values.[49] Labrecque mentions the "crisis of touch" in healthcare AI will likely exacerbate, as it will potentially replace human caregivers.[50] Checketts considers the effects of certain uses of AI on the poor.[51] Scherz calls out the negative quality of power concentration.[52] Vicini likewise is concerned by uses of AI that threaten social justice.[53] Slattery is particularly concerned with the interface of technology and society, where he sees the impact of AI likely to falter, go astray, and cause injustice.[54] Not the being of the technology itself or even its use are in question, but more specifically the use-in-context, most relevant to determining the ethical situation and proper response.

If use is action, and actions come forth from being, then of course being and action can only be intimately linked. Placing the emphasis

---

[46] Graves, "Theological Foundations for Moral Artificial Intelligence," 182–211.

[47] Green, Gaudet, Checketts, Cutter, Herzfeld, Labrecque, Ramelow, Scherz, Vega, Vicini, and Wales, "Artificial Intelligence and Moral Theology," 13–40.

[48] Green, "The Vatican and Artificial Intelligence," 212–31.

[49] Noreen Herzfeld, "Can Lethal Autonomous Weapons be Just?" *Journal of Moral Theology* 11, special issue 1 (2022): 70–86 and Green, Gaudet, Checketts, Cutter, Herzfeld, Labrecque, Ramelow, Scherz, Vega, Vicini, and Wales, "Artificial Intelligence and Moral Theology," 13–40.

[50] Green, Gaudet, Checketts, Cutter, Herzfeld, Labrecque, Ramelow, Scherz, Vega, Vicini, and Wales, "Artificial Intelligence and Moral Theology," 13–40.

[51] Levi Checketts, "Artificial Intelligence and the Marginalization of the Poor," 87–111 and Green, Gaudet, Checketts, Cutter, Herzfeld, Labrecque, Ramelow, Scherz, Vega, Vicini, and Wales, "Artificial Intelligence and Moral Theology," 13–40.

[52] Green, Gaudet, Checketts, Cutter, Herzfeld, Labrecque, Ramelow, Scherz, Vega, Vicini, and Wales, "Artificial Intelligence and Moral Theology," 13–40.

[53] Andrea Vicini, SJ, "Artificial Intelligence and Social Control: Ethical Issues and Theological Resources," *Journal of Moral Theology* 11, special issue 1 (2022): 41–69 and Green, Gaudet, Checketts, Cutter, Herzfeld, Labrecque, Ramelow, Scherz, Vega, Vicini, and Wales, "Artificial Intelligence and Moral Theology," 13–40.

[54] John Slattery, "We Must Find a Stronger Theological Voice: A Copeland Dialectic to Address Racism, Bias, and Inequity in Technology," *Journal of Moral Theology* 11, special issue 1 (2022): 112–31.

on one side or another changes how we look at the ethics of AI. While wanting "good AI" in terms of being seems a worthy goal, it might be enough to aim for "good uses" of AI. There is a whole additional level of inquiry as to whether AI even has being or not, or whether it—as a tool made by human minds and purposes—only has uses.

In his paper, Wales likewise plays at this boundary, wondering whether only the uses or the being of algorithms are subject to ethical inquiry. Opting neither solely for AI use nor being, he most importantly argues that *human being* matters with respect to AI. Wales concludes:

> The right use of AI does not depend merely on the architecture of our systems, nor even on the ethics that we attempt to embed in them, but on the ultimate stance of will that we adopt —be it *superbia* or *caritas*, unto a false knowledge or a true *scientia* and, finally, wisdom. This is the challenge of AI, our moral framing of which will determine what of reality we permit ourselves to see.[55]

This Augustinian interpretation of AI focuses more on how we are involved because AI acts to reflect God's creation back to us.[56]

AI can be used for good or bad things. If endowed with an intrinsic ontology, it might be called good or bad, depending on those consistent dispositions towards action. AI is created by people, for people, to affect people, and therefore ultimately the ethical question resides with us: what kinds of people will be creating and using AI? If we are technically and/or morally bad, we will create technically and/or morally bad AI that takes technically and/or morally bad actions. If instead we follow the better angels of our nature, we might create AI that is both technically and morally better—not perfect.

Being Forced from Participants into Spectators: AI as a Centralizing Disempowering Force

In his paper, Dell'Oro wonders whether AI can be a person and concludes it cannot. AI lacks critical attributes necessary for personhood, including agency and openness to the other.[57] In AI there never really is a participant in activities, only the human-made delegation of participation from someone else. Nor can AI even really observe; observation is delegated by humans and handed back to them. Human beings are not like this—at least not *meant* to live like this.

The gap between participation and observation is a significant one, deserving of more thoughtful consideration. I will raise one example: the Covid-19 pandemic has rendered this gap rather apparent when it

[55] Wales, "Metaphysics, Meaning, and Morality," 181.
[56] Wales, "Metaphysics, Meaning, and Morality," 157–81.
[57] Dell'Oro, "Can a Robot Be a Person?," 132–56.

comes to attending church. A livestreamed video of a church service really is no substitute for being there in the flesh, embodying a community of believers. Screens turn us into observers, and this observational quality has, in some cases, followed us to in-person services as well, as we wonder how to act in-person again, as participants.[58]

It should not be missed that this is a general quality of screens versus real-life. In real-life we participate, on screens we observe. We are not part of the activity of life; we are reduced to viewing the activity of others.

There are huge benefits to viewing church through screens if the alternative is no participation at all. The Beatific Vision might be worth remembering here: someday, God willing, we will see God in-person, united. We should not forget that we are called to be the Body of Christ on earth right now, and through the Eucharist we become what we eat. Participation in the life of God right now enables greater participation in the life of God in the future.

Artificial intelligence could be used right now for better things. AI could assemble all the texts of faith into one *Ultimate General Art*—as Raymond Llull once aspired to produce—which could help form us in the ways of our saints and ancestors. This is a real possibility—but who will do the work? This education and training towards the good aids the formation of souls towards God. We can do this as individuals in community with each other as guides, but AI here makes our community much larger, to encompass anyone who left cultural traces of benefit to the community. This "observation" of our community and identity extended through time does not end with mere observation either. Its end goal is action: being the people of God on this earth, now. Existence is action. We participate in this action, and while observation can prepare us for action, it cannot in itself be our ultimate goal in this life.

As Bishop Tighe notes, there is a gap between the opportunity for life-giving uses of AI and what we have now.[59] AI algorithms which auto-play addictive content shift us from being active participants in life into being deactivated observers of life. We become individuals who live for others' ends—not in a charitable, generous, life-giving way, but a greedy, enslaving, life-taking sort of way—taking hold of our time with the algorithm as instrument of subjugation. This vampirism converts God's concrete and particular gift of innumerable individual lives into the abstraction of money. Corporations which greedily demand more eyeballs viewing their content are stealing the lives of their users in order to turn that life into a resource to empower themselves.

---

[58] I am not arguing against reasonable public health restrictions on community gathering, merely noting that this shift is a significant one.

[59] Green, "The Vatican and Artificial Intelligence," 212–31.

Christianity calls us to participation, not mere observation. Being a spectator of life is not enough, we must actually live. In the context of AI, we should be aware of AI's disempowering tendency and its ability to change us from participants into mere spectators of life. Democracy requires participation, not mere observation. Ethics requires participation, not mere observation. Unjust centralizing powers demand we become mere observers of the collective life administered centrally by the state (really a small group of humans who illegitimately set themselves above others) and this is why Christianity fundamentally is opposed to authoritarianism and totalitarianism. Such critique also applies to unjust corporate and economic centralization. God made us to live, not merely observe.

As exploitative AI tries to change us from living participants in the glorious creation of God into observers subjugated by parasitic others, let us choose the better path. Let us continue to live, observing and even more so acting when appropriate to preserve our agency and use it for good. AI which trains our attention and promotes activity can help in the resistance against unjust uses of AI.

Artificial intelligence is a tool and every tool exists for a purpose. While past tools were much more specific in their aims, intelligence is itself the maker of tools, capable of transforming mere thought into beautiful or brutal reality. AI will be that too: it will merely allow us to get what we want, more of it, faster, more intensely than ever before.

In a context where we are empowered to get anything we want, with little regard to the consequences, wanting the right things becomes of paramount importance. Desire becomes the ultimate power that must be controlled. While Hans Jonas spoke of the much coveted "power over power,"[60] we must now speak of the much needed "desire for desire"—specifically the desire not only for the good, but the best. We must become holy as God is holy, or else we will become dead as sin is dead, as all contingent and evil things must become.[61]

This is no prescription, only a description: contingent beings not purely good and powerful enough to destroy themselves *must* at some point destroy themselves just due to stochasticism. Extinction is only a matter of time, unless we turn towards God, and/or relinquish those powers which threaten to destroy us.[62] We need to reject even the desire for this evil—see, e.g., *Pacem in Terris*, no. 113: "Unless this process of disarmament be thoroughgoing and complete, and reach men's very souls, it is impossible to stop the arms race."

---

[60] Jonas, *The Imperative of Responsibility*, 141–42.

[61] Green, "The Technology of Holiness," 223–28.

[62] Bill Joy, "Why the Future Doesn't Need Us," *Wired*, April 1, 2000, archive.wired.com/wired/archive/8.04/joy_pr.html.

Powerful Technology, Clear Mortality, Ethical Deficiency: Are We, Then, Doomed?

Reverend Martin Luther King, Jr. once lamented that we live in a nation of "guided missiles and misguided men."[63] The situation has not improved over the last 50 years, it may in fact have degraded. No amount of mere intelligence, artificial or otherwise, can make us ethical. A grander vision of intelligence—one including wisdom, flourishing, and holiness—could, but not the impoverished idea of intelligence as "problem solving" or "achieving goals" AI theorists offer us today.[64] Mere "problem solving" without the wisdom of solving the right problems will merely accelerate our decline. We will become very efficient at everything we do, both good and evil. Because it is easier to destroy than create, this asymmetry can only have a sad ending.

In order for AI to be ethical we human beings have to be ethical, and that is difficult. For thousands of years individuals have aimed at holiness, and while we recognize saints, we also recognize that the vast majority of us fall short of that lofty category. Even saints are not perfect, but we are called to try.

In a 2019 lecture to Nobel Laureates, Turing Award winning cryptologist Martin Hellman recalled one of his mentors, business law professor Harry Rathbun, commenting on the question of whether we are "doomed." Hellman said:

> Harry pointed out that there are two hypotheses: Either we are capable of the great changes needed to ensure humanity's survival—that's the nobler hypothesis—or we are not. If we assume the less noble hypothesis, we will be doomed even if we have the capacity to change. But, if we assume the nobler hypothesis, the worst that happens is we go down fighting. And the best that happens is that humanity continues its awesome evolutionary arc. Why not assume the nobler hypothesis?[65]

---

[63] Martin Luther King, Jr., "The Man Who Was a Fool," in *Strength to Love* (Minneapolis, MN: Fortress, 2010).

[64] For just two examples of this kind of rhetoric, see Max Tegmark, *Life 3.0* (New York: Knopf, 2017), 50: "Intelligence = ability to accomplish complex goals"; Tom Simonite, "How Google Plans to Solve Artificial Intelligence," *MIT Technology Review*, March 31, 2016, www.technologyreview.com/2016/03/31/161234/how-google-plans-to-solve-artificial-intelligence/ where DeepMind's Demis Hassabis says he is "solving intelligence, and then using that to solve everything else" (thus endorsing the "intelligence is problem solving" paradigm).

[65] Martin E. Hellman, "The Technological Imperative for Ethical Evolution," Heidelberg Lecture, Lindau Meeting of Nobel Laureates, July 3, 2019, www.mediatheque.lindau-nobel.org/videos/38240/2019-meeting-heidelberg-lecture-hellman. Portions of the speech adapted from Dorothie and Martin Hellman, *A New Map for Relationships: Creating True Love at Home & Peace on the Planet* (USA: New Map, 2016), ee.stanford.edu/~hellman/publications/book3.pdf.

Rathbun and Hellman are right: we should embrace the nobler hypothesis. It is not vain hope: if we believe in God, God gives us that hope.

Let us not be mere observers of a flailing society. Let us be heroic in whatever ways we can be, small or great. God calls us to be like God in holiness. The saints have gone before us and done their part. It is time for us to do ours. We can believe that no matter what happens, God is here and, thankfully, the greatest intelligence.

CONCLUSION

One of my mentors, physician and bioethicist William Hurlbut, once told me: "Always go for the deeper question." AI offers many of these deep questions to pursue, some going straight into the nature of what it means to be human and the fundamental questions of existence, reality, and God. I do not think we yet know the paths that lie before us regarding AI, or the future more broadly. The paths are tangled like yarn, and they only unwind before us as we tread them.

I do know, however, that Christian theologians and ethicists have a special contribution to make. Secular philosophy is not well equipped to deal with some of the fundamentally religious questions raised by AI. Of those secular folks who "get it," some might do brilliant philosophy, while others might do strange-seeming things, like starting churches to AI. Either way, Christian theologians need to contribute what they know and understand. Whether at the center or at the margins, we can enrich this conversation.

In the introduction, my co-editor Matthew Gaudet proposed, for this volume, the metaphor of the hourglass: we narrow the enormous world of AI down into a few topics we can engage in this limited space, but then broaden it again at the end, to recover that wider perspective once more.[66] Squeezing a vast subject into a few papers is not easy and certainly not fair to the field—always biased by selection, no scholarship is fair in this sense. Only by expanding and diversifying the work and workers to truly investigate a field over a long period of time can we hope to do justice to a subject.

For this epilogue I chose the metaphor of threads being drawn together rather messily. There are also other metaphors at play: God as AI, paths to explore, levels of understanding, etc. To be explicit: we ought to understand that the use of the word "intelligence" in AI also is a metaphor and that indeed every comparison of human, machine, and God is metaphorical in some way too. When dealing with difficult subjects, metaphors can help, but also hinder. We need to beware of them. The whole metaphor of artificial or machine "intelligence" can do more harm than good. It would be preferable to be more literal here

[66] Matthew J. Gaudet, "An Introduction to the Ethics of Artificial Intelligence," *Journal of Moral Theology* 11, special 1 (2022): 1–12.

about our machines; to remember they are our tools, and we their wielders.

This volume has been a wonderful challenge, impossible to achieve without the authors, editors at *JMT*, and especially my colleague and friend Matt. The task of engaging moral theology with technology remains impossible to address without you, the reader, taking it (and switching metaphors) to the next level, and further levels beyond. I am so thankful for the voices heard in this volume, yet very conscious that many are missing, which we are in great need of hearing. I hope this special issue will be heard as an invitation to all interested in taking the conversation on AI to greater ends. M

Brian Patrick Green is Director of Technology Ethics at the Markkula Center for applied Ethics at Santa Clara University.

Articles available to view
or download at:

jmt.scholasticahq.com

THE JOURNAL OF MORAL THEOLOGY IS SPONSORED BY:
MOUNT ST. MARY'S UNIVERSITY

THE SCHOOL OF ARTS, HUMANITIES, AND SOCIAL SCIENCES
SAINT VINCENT COLLEGE