Distribution of Moral Accountability in connection with AI systems: A Philosophical Reflection

Seminar on the Philosophy of Technology and Religion – Spring 2021

Reynaldo Belfort-Pierrilus, S.J.

May 16th, 2021

In recent years, the term "AI revolution"[1] has been coined to describe a new era in which the development of Artificial Intelligence (AI) algorithms has become the center of attention in the public and private industry. The promising benefits in AI algorithms such as sophisticated data analysis can be of great benefit in many fields including in healthcare, the automobile industry, and social media platforms. However, a central implication that lies behind AI systems is their self-determination ability which can bring questions around moral accountability when it comes to product disasters or undesired social phenomena involving AI. In tragedies such as a car crash caused by a miscalculation of its AI-based, self-driving feature, it can be ambiguous as to who (or whom) holds the greatest moral weight when advanced AI systems are involved.

In this paper, I proceed to take a closer look into one of the novel AI systems that have been under active development in the past decades and its social impact from a philosophical perspective. More specifically, I will center on the following central question: *Given the increasing prominence of Self-Supervised Learning Artificial Intelligence systems, what questions does the development of this technology raises about moral accountability in cases of tragedy linked to it?* I believe that a more nuanced understanding of moral accountability will be required as AI systems

[1] See Gurnani, CP. n.d. *The AI revolution is here. It's up to businesses to prepare workers for it.* Edited by CNN. Accessed May 16, 2021. https://www.cnn.com/2019/05/30/perspectives/ai-business-jobs; Smith, Craig S. n.d. *A.I. Here, There, Everywhere.* Edited by New York Times. Accessed May 16, 2021. https://www.nytimes.com/2021/02/23/technology/ai-innovation-privacy-seniors-education.html.

are increasingly being incorporated in many consumer products. I further believe that *dialog* as a crucial step in the process of technological development currently being exercised in the 21st century must bring together experts in the applied sciences and technology, sociology, and religion as each of these fields plays a fundamental role in shaping how we relate to one another with the planet. Lastly, I present some works for how such dialog may be achieved and how Catholic theology can make a significant contribution in the process.

Introduction

While there is not a single agreed definition, broadly speaking Artificial Intelligence (AI) is a field of study whose main concern is to understand human intelligence and, based on that understanding, build machines that can operate intelligently. A machine is said to have artificial intelligence if the machine can interpret data collected from its environment, learn from it, and take actions in ways that helps it achieve its goals[2]. Because AI is such an immense field, it will be good to specify what kind of AI technology will be the focus of this paper.

Within the field of AI, the sub-field of machine learning has been the center of attention among researchers and technology developers for the past decades. In this sub-field, there are three main areas of active research, all associated with the process of learning: Reinforcement Learning, Supervised Learning, and Unsupervised Learning[3]. Supervised learning (SL) consists in training an AI system[4] by first feeding to it a vast amount of data labeled by humans so that the machine

[2] Based on Russel & Norvig 2009 and Wikipedia's article "Artificial Intelligence". The field of AI is an immense and rapidly increasing field. To keep track of all the topics involve with the field, refer to Russell, Stuart, and Norvig, Peter. *Artificial Intelligence: a Modern Approach, EBook, Global Edition*. Harlow: Pearson Education, Limited, 2016. Accessed May 16, 2021. ProQuest Ebook Central; Wikipedia, ed. n.d. *Artificial intelligence.* Accessed May 16, 2021. https://en.wikipedia.org/wiki/Artificial_intelligence.

[3] Yann LeCun, "Self-Supervised Learning" (presentation, Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), Hilton New York Midtown, New York, New York, USA, February 10, 2020). Accessed May 16, 2021. https://www.youtube.com/watch?v=UX8OubxsY8w&t=2165s.

[4] I will use AI system and AI algorithm interchangeably throughout the paper.

can predict or classify new data on its own. This particular type of learning mechanism is used in popular technologies such as image recognition, natural language processing, content filtering, among others[5]. In reinforcement learning (RL), training the AI system is performed at scalar level; the model receives a single numerical value as reward or punishment for its actions. Then it uses this sequence of rewards and punishments as a strategy to solve problems in a given setting[6]. However, unsupervised learning, or nowadays more commonly referred to as Self-Supervised Learning (SSL) is the AI technology that is getting more prominence because of its potential benefits that may outweigh the benefits of SL or RL. The central idea behind SSL consists of training an AI machine in such a way that the machine can develop a general understanding of the world around it without requiring constant human supervision in its learning process and with minimal training data.

Yann LeCun, who is a Turing Award recipient and one of the major researchers in the field of machine learning, uses the example of children learning in their early stages of growth to illustrate the idea behind SSL AI systems. In his example, he states that by showing a few images of cows to small children, they eventually are able to recognize any cow they see[7]. This kind of learning can be referred to simply as *common-sense* learning. The kind of learning where we humans learn by trial and error, improving in each iteration as we make mistakes in order to develop skills we need to solve certain situations. According to LeCun along with research scientist Ishan Misra, common-sense learning is what is at the core of biological intelligence in both humans and animals, and attempting machines to do the same is what has remained an open challenge of

[5] Yann LeCun, "Self-Supervised Learning" (presentation, AAAI-20, February 10, 2020).
[6] Yann LeCun, "Self-Supervised Learning" (presentation, AAAI-20, February 10, 2020).
[7] LeCunn, Yann, and Ishan Misra. 2021. Self-supervised learning: The dark matter of intelligence. Facebook AI. March 4. Accessed May 16, 2021. https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

AI research since its inception[8]. As a result, LeCun believes that SSL based AI Systems is one of the most prominent ways to approximate such common-sense learning technique. It would be a major step towards building machines with human-level intelligence. While there has been some progress in the development of this kind of technology, there is still much more to go, which is why LeCun and his team are encouraging others to join the effort of accelerating the development of it[9]. It is this kind of technology (SSL AI systems) that will be the focus for the rest of this paper.

What causes excitement around this SSL technology is the fact that it does not require a vast amount of data in its training process. It is a huge benefit compared to many AI systems based on SL, which requires a vast amount of data labeled by humans to be trained[10]. Moreover, it would help address some of the major challenges society faces today, such as proactive detection of hate speech in social media platforms, major improvements for auto-pilot features in automobiles, improvements in computer vision, applications in health and medicine, apart from the potential to significantly outperform already existing AI technologies such as image recognition or natural language processing[11].

However, developing an SSL-based AI system also implies that such a system would have a strong capacity for self-determination. That is, being able to decide on its own based on its continuous self-learning. It is by reflecting on this *self-determination ability* that I then began to ponder the following question: *Would SSL-based AI systems open the door for us humans to,*

[8] LeCunn, Yann, and Ishan Misra. 2021
[9] Yann LeCun, "Self-Supervised Learning" (presentation, AAAI-20, February 10, 2020); LeCunn, Yann, and Ishan Misra. 2021.
[10] Yann LeCun, "Self-Supervised Learning" (presentation, AAAI-20, February 10, 2020);
[11] Goyal, Priya, Armand Joulin, Vittorio Caggiano, and Piotr Bojanowski. 2021. *SEER: The start of a more powerful, flexible, and accessible era for computer vision.* Facebook AI. March 4. Accessed May 16, 2021. https://ai.facebook.com/blog/seer-the-start-of-a-more-powerful-flexible-and-accessible-era-for-computer-vision.

intentionally or unintentionally, deflect moral responsibility in cases of disaster caused in part or in full by it? Such is the question I now investigate in the following sections.

A more nuanced understanding of scenarios involving AI

To illustrate more concretely what I mean by the risk of deflection of moral responsibility in connection with SSL AI systems, we can consider a hypothetical example of a car crash accident that was a result of a miscalculation in some self-driving feature in the car. This example is in part inspired by the car crashes in recent years involving self-driving features such as Tesla's prominent "Autopilot" feature which is powered by AI[12]. Many drivers and passengers have died or suffered serious injuries in connection with some malfunction of this "Autopilot" feature[13].

Let us say that a driver along with a friend passenger has turned on the self-driving feature in the car while driving on a given road in the suburbs of a city. Both the driver and the passenger *trust* the self-driving feature given how well it has worked in the past for them and others. Then the car crashes against an obstacle that could have been easily avoided if the driver and the passenger were supervising the car feature. Both the driver and the passenger die in the crash. *Who is morally responsible for this crash?* One could respond with the following argument: (1) *it was, ultimately, both the driver and the passenger's fault for not supervising the self-driving feature that led to the car crash.* This argument is based on the fact that the humans in the car *trusted too much* the AI system. Putting too much trust into AI systems is a topic of active discussion and

[12] Tesla. n.d. *Autopilot.* Tesla Motors, Inc. Accessed May 16, 2021. https://www.tesla.com/autopilotAI.
[13] See Boudette, Neal E. 2021. *Tesla's Autopilot Technology Faces Fresh Scrutiny.* New York Times, March 23. Accessed May 16, 2021. https://www.nytimes.com/2021/03/23/business/teslas-autopilot-safety-investigations.html. ANTCZAK, JOHN, and TOM KRISHER. 2021. *Crash, arrest draw more scrutiny of Tesla Autopilot system.* ABC News, May 12. Accessed May 16, 2021. https://abcnews.go.com/Weird/wireStory/police-california-tesla-driver-riding-backseat-arrested-77646289;

discussing it here will exceed the scope of this paper[14]. While argument (1) could work, I am afraid that the answer to this question is much more complicated. For this reason, Actor-Network Theory can serve as a great philosophical tool to help us have a more nuanced understanding of the situation.

Actor-Network Theory (ANT), developed by philosopher and sociologist Bruno Latour, along with many others such as John Law and Michel Canon is an increasingly influential yet strongly debated philosophical approach that seeks to understand human behaviors and their interaction with inanimate objects[15]. Broadly speaking, this philosophical approach seeks to understand social processes that emerge as a result of an overall agency that is driven by a network of whose nodes are both humans and inanimate objects. While it is not possible to cover all the complex features of this theory in this paper, I believe its basic concepts can be a great analytical tool for understanding the complexities behind situations such as the self-driving car crash example.

ANT's main feature is laying out the composition of a *network* that is composed of actors. An *actor* is defined as a "source of an action regardless of its status as a human or non-human". What is key about this definition is the radical notion that non-living objects can also have agency. An actor however can only act in combination with other actors and in constellations that give the actor the possibility to act[16]. Based on ANT's own epistemological and ontological position, an

---

[14] For a closer look in the topic of trust in AI Systems, see Hao, Karen. 2021. *We need to design distrust into AI systems to make them safer.* MIT Technology Review, May 13. Accessed May 16, 2021. https://www.technologyreview.com/2021/05/13/1024874/ai-ayanna-howard-trust-robots/?truid=&utm_source=the_algorithm&utm_medium=email&utm_campaign=the_algorithm.unpaid.engagement&utm_content=05-14-2021.

[15] Cresswell, Kathrin M, Allison Worth, and Aziz Sheikh. 2010. *"Actor-Network Theory and its role in understanding the implementation of information technology developments in healthcare."* BMC Medical Informatics and Decision Making volume 10. doi:https://doi.org/10.1186/1472-6947-10-67, 1.

[16] Cresswell, Kathrin M, Allison Worth, and Aziz Sheikh. 2010, 2.

actor can include humans, things, ideas, or concepts. However, for this paper, we will restrict ourselves between humans and *technological objects*. A breath of examples can be considered as a technological object; ranging from hand tools (i.e. hammer) or hardware (i.e. a vehicle brake system) to software such as AI systems or social media platforms. Tracing associations or relationships between actors in a network and how this network evolves over time is a key activity in ANT[17]. However, the composition of this network tends to become of particular interest when things in a system go wrong[18], which is why we will consider ANT in our self-driving car crash example.

Returning to our example, let us then identify the actors involved in the *relevant network* that led to the unintended consequence: the vehicle's crash and the death of the driver and the passenger in it. I say *relevant network* because while ANT theory could be applied to sketch how a particular object is related to virtually anything, our concern here is to trace all the actors that could be considered as morally responsible for the situation. Let us say then that the actors of this simplified relevant network are: the driver, the passenger, the AI system, and the private company that designed the AI system, which can be further composed of a team of software designers and the bosses who supervise them. Notice here that the AI system can now be considered as a relevant actor of this network since it possesses a self-determination ability. Who then should be, ultimately, morally responsible for this car crash? The driver alone? The driver and the passenger? The AI system? Or a combination of all the actors mentioned above plus the software designers, all the way to the bosses of the private company? Again, while this is a simplified sketch of our relevant network, we can see here how the notion of *distributed moral responsibility* comes at

[17] Cresswell, Kathrin M, Allison Worth, and Aziz Sheikh. 2010, 2.
[18] Cresswell, Kathrin M, Allison Worth, and Aziz Sheikh. 2010, 3.

play. The kind of distribution where not only humans are solely involved, but also AI technologies that can decide for themselves; a technology made by humans.

Therefore, some may begin to argue that because AI systems can decide for themselves, human beings (such as the designer or even the driver in our car crash example) should not hold the greatest moral weight when it comes to accountability. But is this true? This is where technology developers such as the designers come into play since, in the end, it is technology developers who are creating AI Systems like the ones in our hypothetical example. This is not to say, however, that technology developers are the ones who should ultimately hold the greatest moral weight when it comes to disasters involving AI. But I do believe that introducing new technologies into the world with some particular intention in mind, can also raise unintended consequences that could be more harmful in the long run. And it seems that we humans delegating responsibility and moral accountability to AI systems such as SSL-based AI systems is perhaps an unintended consequence that can lead us to many social issues[19].

Why developing advanced AI Systems such as SSL-based AI may be of concern?

While our self-driving car crash example was a simplified and hypothetical one, the issue of moral responsibility extended to AI Systems may also show up in many other cases that perhaps are much more delicate. Especially concerning technologies that can significantly influence social behavior such as social media platforms (SMPs). For instance, one of the promising benefits of SSL AI Systems is the proactive detection of hate speech (SMPs) such as Facebook, Twitter or Instagram, etc.[20] Which would be a good thing. Yet it is also known that AI Systems forms the

[19] Whether delegating moral accountability is intended or unintended by technological developers can be left for debate. Either way, it is a topic that the human community should take into serious consideration.
[20] Yann LeCun, "Self-Supervised Learning" (presentation, AAAI-20, February 10, 2020); LeCunn, Yann, and Ishan Misra. 2021.

bulk of recommendation algorithms and personalized advertisement algorithms (i.e. YouTube), or any SMP[21]. Moreover, studies are being conducted on how SMPs (many of which strongly depend on AI systems[22]) are impacting people (who actively engage in these platforms) in understanding social issues centered more on perception rather than objective reality[23]. There is also constant debate into who is to be ultimately held accountable for the spread of misinformation and extremism involving SMPs[24]. Many AI-based software systems are also under constant scrutiny and public concern such as the use of facial recognition in Law Enforcement[25], hiring bias in private companies[26], the pursuit of Artificial General Intelligence[27], and concern on the environmental impacts tied to product design processes that involve AI to some capacity[28]. All of the cases mentioned above can present a form of the moral accountability problem illustrated in our self-driving car example. At the same time, it is worth mentioning as well that, sophisticated

[21] Some documentaries explore how these algorithmic features are at play in different web applications. For more a closer look into this topic, see 2019. *The Great Hack.* Directed by Karim Amer and Jehane Noujaim. Netflix; 2020. *The Social Dilemma.* Directed by Jeff Orlowski. Netflix.

[22] Yann LeCun, "Self-Supervised Learning" (presentation, AAAI-20, February 10, 2020);

[23] 2019. *The Great Hack.* Directed by Karim Amer and Jehane Noujaim. Netflix; Kaufmann, Eric. 2021. *The Social Construction of Racism in the United States*. Manhattan Institute for Policy Research, Inc. April 7. Accessed May 16, 2021. https://www.manhattan-institute.org/social-construction-racism-united-states.

[24] 2021. *Joint Hearing: "Disinformation Nation: Social Media's Role in Promoting Extremism and Misinformation" - Witnesses: Mr. Mark Zuckerberg, Mr. Sundar Pichai, Mr. Jack Dorsey*. Washington, D.C.: U.S. House of Representatives Committee Repository, March 25. Accessed May 16, 2021. https://docs.house.gov/Committee/Calendar/ByEvent.aspx?EventID=111407.

[25] Mac, Ryan, Caroline Haskins, and Logan McDonald. 2020. *Clearview's Facial Recognition App Has Been Used By The Justice Department, ICE, Macy's, Walmart, And The NBA.* BuzzFeed News, February 27. Accessed May 21, 2021. https://www.buzzfeednews.com/article/ryanmac/clearview-ai-fbi-ice-global-law-enforcement.

[26] For one particular example, see Dastin, Jeffrey. 2018. *Amazon scraps secret AI recruiting tool that showed bias against women.* Reuters, October 10. Accessed May 16, 2021. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[27] Hao, Karen. 2020. *The messy, secretive reality behind OpenAI's bid to save the world.* MIT Technology Review, February 17. https://www.technologyreview.com/2020/02/17/844721/ai-openai-moonshot-elon-musk-sam-altman-greg-brockman-messy-secretive-reality/.

[28] For a deeper look into the impact of AI in the environment, see Dryer, Theodora. February 3rd, 2021. *Testimony before the European Parliament Greens/EFA Group. A Digital and Green Transition Series: Will Artificial Intelligence Foster or Hamper the Green New Deal?* Republished on Medium by the AI Now Institute at New York University; Crawford, Kate, and Vladan Joler. 2018. *Anatomy of an AI System: The Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources.* AI Now Institute and Share Lab, September 7. Accessed May 16, 2021. https://anatomyof.ai.

AI systems like SSL algorithms could become the next-level AI technology that will replace existing and more limited AI Systems.

What does the Catholic Church can contribute to the question of responsible technological development?

After discussing the most recent and promising AI technology to this day (SSL-based AI), its self-determination ability, and the potential consequences tied to moral responsibility, it is becoming clear that we need to put greater focus into the process of technological development driven by technology developers, especially in the software industry. To have a better understanding of the current process of technological development in the 21st century, the work of Luis O. Jiménez-Rodríguez, S.J. can be of help.

Jiménez-Rodríguez defines *technocratic instrumentalism* as a generic ideology that is rooted in fundamental beliefs, an implicit anthropology, a conception and vision of reason, a subjacent philosophy of nature, and a moral imperative[29]. Jiménez-Rodríguez states the following to be the **fundamental beliefs** within technocratic instrumentalism: (1) Technology is morally neutral and is the drug that will solve all human, social and environmental problems. (2) Natural resources exist in unlimited quantity and to our disposition. (3) Progress is indefinite[30]. For the **implicit anthropology**, it is the vision that conceives the human being as a subject which end is control, hegemony, and transformation of nature with the objective of taking total advantage of its resources to develop consumer goods[31]. Jiménez-Rodríguez further states that related to this

[29] Jiménez-Rodríguez SJ, Luis O. 2019. «Los Aportes De La teología De La creación Y De La acción Humana a La orientación De Las Ciencias Aplicadas Y Las tecnologías: Una mediación ética Y axiológica». *Pensamiento. Revista De Investigación E Información Filosófica* 75 (283 S.Esp), 389. https://doi.org/10.14422/pen.v75.i283.y2019.021.
[30] Jiménez-Rodríguez SJ, Luis O. 2019, 389
[31] Jiménez-Rodríguez SJ, Luis O. 2019, 389

implicit anthropology exists also a **vision of reason** as a rationality reduced to the cognitive instrumental and that puts aside other aspects of reason such as reflection as wisdom about the ultimate goals of personal and social life, prudence or deliberation about human praxis, and hermeneutics about the meaning of life[32]. For the **subjacent philosophy of nature**, technocratic instrumentalism conceives and represents nature as an aggregate of things to the image and likeness to the machines that must be in service of human beings[33]. Finally, technocratic instrumentalism is guided by the following moral imperative: *what can be designed, built or manufactured, must be done*[34].

There is much more to be said about the ideology of technocratic instrumentalism, which can be explored further in his work. But considering this ideology and the key concepts behind it gives us a good idea of what the orientation of technological development consists in our present-day. If the present-day process of technological development remains in its current orientation, we will find ourselves developing an unsustainable world that will harm both the present generation and future generations. For this reason, we must include *dialog* as an additional step in our present process of technological development. But the kind of dialog that brings both experts in the fields of applied sciences and technologies and social experts that can capture and bring in the voices of people, especially the poor and the marginalized. Moreover, special consideration should be given to those who understand the co-evolution of technology and religion throughout history from the slow invention of writing and the Axial Revolution to the Printing Press and beyond, since religion

[32] Jiménez-Rodríguez SJ, Luis O. 2019, 389
[33] Jiménez-Rodríguez SJ, Luis O. 2019, 390
[34] Jiménez-Rodríguez SJ, Luis O. 2019, 390

plays a major role in society[35]. This can help us anticipate unintended consequences that can emerge from newly created technologies.

As for how this dialog could take place, Jimenez-Rodríguez in his work proposes the Ethics and Axiology as a mediation between the contributions of Catholic theology (theology of creation and co-creative human action) to the re-orientation of technological development, and the techno-sciences[36]; where deliberation will serve as the key activity for this mediation process[37]. Taking a closer look into this proposition will not be possible in this paper, but the reader is encouraged to examine his work. Nevertheless, this proposal opens the door for a dialog that can form a bridge between the ecumenical (world religions) and the secular.

In conclusion and to be clear, I do not argue or advocate for discontinuing the development of AI Systems based on SSL or any other technologies. What I argue though is that, given the intended and unintended consequences that AI is presenting (especially SSL-based AI), we need "all hands on deck" to further examine its impact on society and to determine whether its development should be continued or altered in some way. The same process of examination can be applied to other controversial technologies such as genetic engineering and so on. It is this human skill that will help us create and maintain a more sustainable social and ecological environment for the present and future generations.

---

[35] The work of Tim Clancy S.J. and Walter Ong, S.J. may be of interest. See S.J., Dr. Tim Clancy. n.d. *Jesuit Seminar on Religion and Technology*. Accessed May 16, 2021. https://religioustech.org/; Walter J. Ong, SJ. 1998. *"Digitization Ancient and Modern: Beginnings of Writing and Today's Computers."* Communication Research Trends (Centre for the Study of Communication and Culture - Saint Louis University) 2: 18.

[36] Jiménez-Rodríguez SJ, Luis O. 2019, 397

[37] Jiménez-Rodríguez SJ, Luis O. 2019, 397-405