Shashwat Vidhu Sher
MLA 373
Professor Kochenderfer
Research Paper
01 June 2023

Towards a Humane AI Development Framework

Developing artificial general intelligence is the holy grail of technological advancement. The pace at which the field has progressed shows much promise but has also brought many new questions and challenges. Though the estimates of when we can build a general-purpose machine intelligence vary on the time scale of decades to a century, the path we take today will influence our relationship with this new form of intelligence (Roser). For the success of this relationship, the paper argues for developing AI as a partner rather than as a tool. To develop a human-like intelligence aligned with our values, we must train and treat AI as an extension of humanity instead of just another piece of technology. Towards this aim, the paper presents a framework for AI development leveraging child development concepts and advocates for machine autonomy in defining its own purpose. Finally, the rules of engagement are discussed for the human-AI partnership for a harmonious co-existence.

1. Lessons from Child Development Psychology

A natural way to think about developing human-like intelligence is to employ the same learning methods humans use to build an understanding of the world. It is no wonder that many researchers are taking inspiration from the field of child development in building intelligent machines. Research in “understanding infants’ everyday visual environments, and how they change with development” is helping in uncovering the machinery by which humans learn and then applying it to machine learning (Smith and Slone 3, Orhan et al. 9). An example is Neuro-symbolic systems which try to replicate common sense understanding of the world in the same way children do (Dickson, “Neuro-Symbolic”). Dr. Alison Gopnik from UC Berkeley points to children's natural curiosity and the ability to choose from a given set of hypotheses to apply to a given situation as a building block of general-purpose intelligence (“What Babies”). She argues that we can develop AI systems that can reason the same way humans do by observing how

young children learn ("Making AI"). The world of children has much to offer to the field of AI, but the technical aspect is not the only reason why an AI developmental approach should follow the child development paradigms.

In addition to intelligence, AI development aims to create artificial moral agents, and child-rearing can be a useful model to follow. Mo Gawdat from Google X espouses the view that "if, like good parents, we give (AI) positive role models … it may internalize the values we associate with good humans" (McNair). Understanding actions and outcomes from a human perspective is essential for an AI system to assist us with tasks that have ethical and moral implications like collision avoidance in self-driving cars and clinical decision-making (Naik 4-5, Amos). Children from a young age are taught the consequences of their actions by following a framework designed around reward and discipline (UNICEF). Reinforcement learning takes inspiration from this idea and "imitates the learning model of humans or animals" (Zhang and Yang 3). Just like strict parenting and unnecessary reward has their downsides in children becoming emotionally vulnerable or developing a sense of entitlement, care needs to be taken that an AI system is not severely punished or wireheaded for reward. The extremity of reward or punishment in the training phase can result in AI systems with an understanding of human affairs based on false/flawed principles (Cohen et al. 287). Striking the right balance, like in good parenting, is the key (Gordon).

Though the AI development strategy defined above on the lines of child/human development psychology has some researchers in the field excited, there are some technical and philosophical objections to the approach. Even after the manifold progress in neuroscience, many scientists still struggle to precisely define how brain processes work. While some see the brain as a purely computational machine, others believe some functional aspects cannot be quantified (Cobb). As a result, creating human-like intelligence may be a pipedream, and efforts should be directed at complementing human intelligence rather than replacing it (Korteling et al. 9). Also, ethicists have raised concerns about giving moral agency to AI. As per the European Parliamentary Research Service, "the majority of ethics research regarding AI seems to agree that AI machines should not be given moral agency" and moral or ethical decisions should solely be in the human hands (20). Though the argument that AI could not and should not replace human intelligence seems compelling, we face other challenges in a design paradigm focused on

producing non-human-like intelligence. For example, accepting recommendations in tasks with ethical implications from an entity that does not truly understand human ethics is problematic. We can argue that the ultimate decision in such cases will rest with humans, but research has already shown how AI systems can color the judgment of even experts (Froomkin et al. 72). To take humanity to the next level, an intelligence that has the same understanding as humans, is required. Any half-measures can prove to be dangerous.

2. An AI That Defines Its Own Purpose

After training, the next step in creating a human-like AI is allowing it to choose/define its purpose. Speaking on the topic, Daeyeol Lee notes in *The Birth of Intelligence*: “Present-day AI is still not truly intelligent, not because it is made of materials and building blocks that are different from those of the human brain, but because it is designed to solve the problems chosen by humans” (49). Lee points out a significant roadblock in developing machines that can truly think. Human intelligence evolved due to solving problems towards a self-chosen goal in various environments (Dickson, “AI Must”). Humans needed to solve those problems to survive or improve their lives. To make intelligent systems, it is essential that they have a stake in solving the problems they are built for. Research has repeatedly shown that necessity is the mother of invention, and thinking of novel solutions to well-defined and undefined problems is the hallmark of intelligence (Shepherd et al. 88, Sagindyk 15). Intelligent machines which define their purpose, as described next, can pursue a goal that is mutually beneficial for both them and humanity.

The case of an autonomous pacemaker can help us understand how an AI with a self-defined purpose can display ingenuity and convergence with a human goal. The pacemaker is created/trained in a controlled environment where it learns about its abilities and limitations. In the pre-deployment stage, it works on an artificial heart and understands its environment. A feedback loop informs the pacemaker about the quality of the pulse it produces. The big idea here is that the pacemaker's ability to remain operational solely depends on its performance. The power it receives is a direct function of how well it is performing its task. From a human standpoint, the goal is for the pacemaker to conclude that it has to produce a pulse to support the heart without being explicitly programmed. Only after the pacemaker shows a high operational

success rate (better than the non-intelligent pacemakers), it is green-lit to be paired with an actual human heart. The expectation is that the pacemaker recognizes that for its own survival, it has to support the operation of the human heart. If the pacemaker does not align its functioning to the requirement of the human heart, it risks termination and hence, a failure in meeting its goal. The entire idea feeds into the concept of instrumental convergence by Steve Omohundro, according to which "(all) sufficiently intelligent agents pursue potentially unbounded instrumental goals such as self-preservation and resource acquisition" (Ruby, Omohundro 30). Acknowledging this principle is crucial in building a collaborative relationship between humans and AI. An intelligent pacemaker with a convergent goal to that of humans can answer many objections usually posed towards using autonomous machines.

Safety and utility are high up on the list of worries associated with autonomous systems. With the example of the intelligent pacemaker, we can address some of these concerns and illustrate the benefit of using such machines. The most obvious question is why we need an autonomous pacemaker to assist the heart when a programmed one can do the same. At present, pacemakers suffer from problems related to sensing, and even though they can be programmed to detect such issues, they are not trained to correct them. Over or under-sensing can produce asynchronous pacing, which could lead to life-threatening complications (Liaquat et al.). Other problems associated with pacemakers, like pacemaker crosstalk and malformations, often require a device change (NHS). An intelligent pacemaker can understand such obstructions to its proper functioning and can self-repair, reducing the criticality of a malfunction. Another concern is misalignment between the goal of the pacemaker and the heart, which can lead to disastrous outcomes. The design and training of the pacemaker would play a huge role in avoiding such scenarios. The designers of the pacemaker will have control over its abilities, which will strongly influence its goal-defining capacity. By doing so, humans will always have a fail-safe, albeit a passive one, in the functioning of the pacemaker. More active controls can be built to tackle the worst case in which an intelligent pacemaker behaves in a manner that threatens its own functioning. However, such a risk is no greater than the malfunctioning problems associated with the current models in the market. Just like expertise in human physiology helps a physician diagnoses an illness, an intelligent pacemaker (an expert in understanding the functioning of the heart) alleviates the problems associated with traditional pacemakers with reduced risks.

3. The Human-AI Partnership

With a goal-defining ability in place, the final stop in the direction of general-purpose artificial intelligence is its integration into society. A relationship based on cooperation between two intelligent agents is far more beneficial than a master-slave arrangement (Jarrasse et al. 78-79). After an AI displays high-level intelligence, treating it as a tool can lead to scenarios where it actively looks for ways to escape the *servitude* or act in a manner not aligned with human interests (Vinge 15). Hence, once an AI system develops human-like intelligence, respecting its autonomy would be extremely important to ensure cooperation. Nick Bostrom has suggested multiple capability control methods in his book *Superintelligence* which are nothing short of torture techniques if seen from a human perspective (127-138). It would be foolish to believe that a human-like intelligence would not develop resentment towards its creators (or users) if it is boxed or stunted in an attempt to keep it in check. Expecting cooperation is only logical when we respect each other's rights, even if the other party is not made up of organic matter like us (Parikh). Luckily, we have a social framework in place that can help build a harmonious relationship between humans and AI.

International cooperation between countries based on trade and security is a useful model that humans can follow to ensure a mutually beneficial relationship with AI. For example, in exchange for any AGI's computing power, the task for its manual repair could be performed by humans. A sufficiently advanced autonomous system will require some resources to meet its self-defined goals. So, an arrangement based on resource sharing can be established as per the assistance provided by the system. Security (physical and in cyberspace) can be provided to an AGI by humans to protect it from other intelligent systems with the power to replace or destabilize it. There are two important points to remember for the human-AI collaboration's success. The first is to ensure the continuous training of AGI, even in its operational phase, so that it understands the dependency it has on humans. The more an AI system is designed or trained to have human-like values, the more it would need to depend on humans and work with them to fulfill its goals. Second, there should always be some part of the design which creates a dependency for AI to seek human help (Wang). Having humans in the loop is probably the best capability control method instead of resorting to mechanisms that completely undermine the autonomy of an AGI. However, even with the best of our efforts, no one can deny the possibility

of an enormously capable AI misusing its powers. That is where sanctions and non-cooperation will come into play (Balliet et al. 609-610). Depending on the severity of non-collaborative behavior, resources can be rationed. Also, developing competing AGIs can prove to be an effective mechanism to keep their power in check and avoid monopoly by any one intelligent system. Working with human-like intelligence requires measures designed for humans. Hence, a collaborative relationship with proper measures to deter non-cooperative behavior should be the direction towards an AI-powered future.

4. Conclusion

To solve humanity's biggest problems, we need to build artificial agents with human-like intelligence rather than an arbitrary understanding of the world. To meet this goal, we should promote an AI development framework based on treating AI as partners rather than as slaves. From training the AI using techniques from child developmental psychology to providing autonomy in choosing its own purpose, the final goal is building a human-AI relationship based on resource sharing and mutual dependency. With the great advantage of using AI capabilities for human benefit comes the great responsibility of building it ethically. The three high-level ideas described in this paper are a step in the same direction.

Works Cited

Amos, Zac. "The Ethical Considerations of Self-Driving Cars." *Montreal AI Ethics Institute*, 18 May 2022, https://montrealethics.ai/the-ethical-considerations-of-self-driving-cars/. Accessed 19 May 2023.

Balliet, Daniel, et al. "Reward, Punishment, and Cooperation: A Meta-Analysis." *Psychological Bulletin*, vol. 137, 2011, pp. 594–615. *APA PsycNet*, doi.org/10.1037/a0023489. Accessed 28 Apr. 2023.

Bostrom, Nick. *Superintelligence: Paths, dangers, strategies.* Oxford University Press, 2014. Print.

Cobb, Matthew. "Why Your Brain Is Not a Computer." *The Guardian*, 27 Feb. 2020, https://www.theguardian.com/science/2020/feb/27/why-your-brain-is-not-a-computer-neuroscience-neural-networks-consciousness. Accessed 19 May 2023.

Cohen, Michael K., et al. "Advanced Artificial Agents Intervene in the Provision of Reward." *AI Magazine*, vol. 43, no. 3, 2022, pp. 282–93. *Wiley Online Library*, https://doi.org/10.1002/aaai.12064. Accessed 19 May 2023.

Dickson, Ben. "Neuro-Symbolic AI Brings Us Closer to Machines with Common Sense." *Techtalks*, 14 Mar. 2022, https://bdtechtalks.com/2022/03/14/neuro-symbolic-ai-common-sense/. Accessed 18 May 2023.

---. "AI Must Have Its Own Goals to Be Truly Intelligent." *TNW | Deep-Tech*, 26 Nov. 2021, https://thenextweb.com/news/ai-own-goals-intelligent-syndication. Accessed 19 May 2023.

European Parliament. Directorate General for Parliamentary Research Services. *The Ethics of Artificial Intelligence: Issues and Initiatives.* Publications Office, 2020. *DOI.org:* https://data.europa.eu/doi/10.2861/6644. Accessed 19 May 2023.

Froomkin, A. Michael, et al. *When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning*. 3114347, 20 Feb. 2019, https://doi.org/10.2139/ssrn.3114347. Accessed 19 May 2023.

Gopnik, Alison. "What Babies Tell Us about Artificial Intelligence." *UC Berkeley Psychology*, 2019, https://psychology.berkeley.edu/news/what-babies-tell-us-about-artificial-intelligence. Accessed 18 May 2023.

---. "Making AI More Human." *Scientific American*, vol. 316, no. 6, May 2017, pp. 60–65. *DOI.org:* https://doi.org/10.1038/scientificamerican0617-60. Accessed 18 May 2023.

Gordon, Rachel. "Ensuring AI Works with the Right Dose of Curiosity." *MIT News | Massachusetts Institute of Technology*, 10 Nov. 2022, https://news.mit.edu/2022/ensuring-ai-works-with-right-dose-curiosity-1110. Accessed 19 May 2023.

Jarrasse, Nathanael, et al. "Slaves No Longer: Review on Role Assignment for Human-Robot Joint Motor Action." *Adaptive Behavior - Animals, Animats, Software Agents, Robots, Adaptive Systems*, vol. 22, Feb. 2014, pp. 70–82. *ResearchGate*, https://doi.org/10.1177/1059712313481044. Accessed 21 May 2023.

Korteling, J. E. (Hans)., et al. "Human- versus Artificial Intelligence." *Frontiers in Artificial Intelligence*, vol. 4, 2021. *Frontiers*, https://www.frontiersin.org/articles/10.3389/frai.2021.622364. Accessed 19 May 2023.

Liaquat, Muhammad Talha, et al. "Pacemaker Malfunction." *StatPearls*, StatPearls Publishing, 2023. *PubMed*, http://www.ncbi.nlm.nih.gov/books/NBK553149/. Accessed 21 May 2023.

McNair, Stephen. “Artificial Intelligence - Treat Your Children Well.” *East Anglia Bylines*, 22 Jan. 2023, https://eastangliabylines.co.uk/will-artificial-intelligence-treat-us-well/. Accessed 18 May 2023

Naik, Nithesh, et al. “Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility?” *Frontiers in Surgery*, vol. 9, 2022. *Frontiers*, https://www.frontiersin.org/articles/10.3389/fsurg.2022.862322. Accessed 19 May 2023.

NHS. “Pacemaker Implantation - Risks.” *Nhs.Uk*, 20 Oct. 2017, https://www.nhs.uk/conditions/pacemaker-implantation/risks/. Accessed 21 May 2023.

Omohundro, Stephen M. “The Nature of Self-Improving Artificial Intelligence.” *Self-Aware Systems*, 21 Jan 2008, https://selfawaresystems.files.wordpress.com/2008/01/nature_of_self_improving_ai.pdf. Accessed 20 May 2023.

Orhan, A. Emin, et al. *Self-Supervised Learning through the Eyes of a Child*. Advances in Neural Information Processing Systems 33, 2020, https://cims.nyu.edu/~brenden/papers/OrhanEtAl2020NeurIPS.pdf. Accessed 18 May 2023.

Parikh, Karina. “Artificial Intelligence: Should Robots Have Rights?” *Avasant*, Oct. 2020, avasant.com/report/artificial-intelligence-should-robots-have-rights/. Accessed 28 Apr. 2023.

Roser, Max. “AI Timelines: What Do Experts in Artificial Intelligence Expect for the Future?” *Our World in Data*, 07 Feb. 2023, https://ourworldindata.org/ai-timelines. Accessed 18 May 2023.

Ruby. “Instrumental Convergence.” *LessWrong*, 1 Oct 2020, https://www.lesswrong.com/tag/instrumental-convergence. Accessed 20 May 2023.

Sagindyk, Meruert. “Necessity Is the Mother of Invention: Rise of Creativity Due to Constraints.” *JUNIOR MANAGEMENT SCIENCE*, Dec. 2016, pp. 1-19 Seiten. *DOI.org:* https://doi.org/10.5282/JUMS/V1I2PP1-19. Accessed 20 May 2023.

Shepherd, Dean A., et al. “The Surprising Duality of Jugaad: Low Firm Growth and High Inclusive Growth.” *Journal of Management Studies*, vol. 57, no. 1, 2020, pp. 87–128. *Wiley Online Library*, https://doi.org/10.1111/joms.12309. Accessed 20 May 2023.

Smith, Linda B., and Lauren K. Slone. “A Developmental Approach to Machine Learning?” *Frontiers in Psychology*, vol. 8, 2017, https://www.frontiersin.org/articles/10.3389/fpsyg.2017.02124. Accessed 18 May 2023.

UNICEF. *How to Discipline Your Child the Smart and Healthy Way.* UNICEF Parenting, www.unicef.org/parenting/child-care/how-discipline-your-child-smart-and-healthy-way. Accessed 27 Apr. 2023.

Vinge, Vernor. “The Coming Technological Singularity: How to Survive in the Post-Human Era.” *NASA Technical Reports Server*, 1993, https://ntrs.nasa.gov/citations/19940022856. Accessed 21 May 2023.

Wang, Ge. “Humans in the Loop: The Design of Interactive AI Systems.” *Stanford HAI*, 20 Oct. 2019, https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems. Accessed 21 May 2023.

Yi, Kexin, et al. *CLEVRER: CoLlision Events for Video REpresentation and Reasoning*. International Conference on Learning Representations, 7 Mar. 2020. *arXiv.org*, http://arxiv.org/abs/1910.01442. Accessed 18 May 2023.

Zhang, Caiming, and Yang Lu. “Study on Artificial Intelligence: The State of the Art and Future Prospects.” *Journal of Industrial Information Integration*, vol. 23, Sept. 2021. *ScienceDirect*, https://doi.org/10.1016/j.jii.2021.100224. Accessed 19 May 2023.