International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Artificial Intention: A Path to Ethical AI?

Pawas Piyush

Institute of Management Studies, Noida, ChaudharyCharan Singh University
Email: *pawasbiz[at]gmail.com*

Abstract: ***This article explores the intricate relationship between artificial intelligence AI and ethical concerns, focusing on the need to understand and control the intentions of AI systems. It delves into the risks associated with AIs lack of moral guidance and its potential to produce damaging content, emphasizing the importance of ethical oversight in the development and deployment of AI technologies. The article also examines current efforts by tech giants to mitigate ethical challenges and classifies these solutions into two key categories: probing and modeling support and restrictions on defined outputs. It introduces the concept of Artificial Intention as a novel frontier in AI ethics and highlights the importance of aligning AI solutions with a planet - centric approach to address pressing global challenges. Ultimately, the article underscores the role of ethics and intentionality in shaping a responsible and equitable future for AI.***

Keywords: Artificial Intelligence, Ethics, Artificial Intention, Ethical Oversight, Planet - Centric, AI Development, Ethical Challenges, Tech Giants, AI Ethics Tools

We are part of a race. In fact, we have always been. And why shouldn't it be this way? Isn't that the survival of the fittest always ingrained in our DNAs? So, if it is a race for survival, then we will most certainly do everything possible to control our environment. Control to make it favourable to us. Favourable to our self - interest, favourable to our businesses, favourable to our environment, and so on and so forth. This is also widely known as ***"The Law of Requisite Variety"***. (1)

Assume we have a remote control for our television with a variety of buttons. Each button indicates a particular function, such as changing the channel or increasing or decreasing the volume. If we want to efficiently control our TV, we will need a remote control with as many buttons as the functionality on our TV. This means that the number of buttons on our remote control should correspond to the number of features on your TV.

Now, consider the environment to be a very complex system comprising numerous components such as ecosystems, weather patterns, and natural processes. Humans often wish to control or manipulate the environment to make it suit their needs or desires, just like using a remote control to change the channel on the television.

The Law of Requisite Variety states that if we wish to effectively control the environment, we must have as many diverse methods (or strategies) as there are different parts of the environment that we want to control. In other words, our "control remote" should contain as many buttons as there are functions in the environment we're attempting to administer or manage.

If we don't have enough methods or "buttons" to match the complexity of the environment, our attempts to control it may fail, just like trying to operate a TV with a remote control that has too few buttons. Therefore, the Law of Requisite Variety reminds us that understanding and managing the environment involves a diverse set of approaches and solutions to be effective.

Now that we understand ***why*** we want to control our environment, let's also try and figure out ***how* to do it**. How do we go about it? Since there are so many variables to consider and control, it is not humanly possible for us to keep track of them all and control or modify them individually. It is here that we take the help of the machines.

Machines have come a long way on their journey from humble beginnings to the realm of artificial intelligence. Starting as simple mechanical devices like the abacus and early calculators, they gradually evolved into complex machines, such as the early computers of the mid - 20th century, which could perform mathematical calculations with remarkable speed. Over time, the integration of transistors, microprocessors, and software programming ushered in the digital age, giving rise to personal computers and the internet. This rapid technological progress paved the way for artificial intelligence, enabling machines to learn, reason, and process information like never before. Machine learning algorithms, big data, and neural networks have all contributed to the development of AI systems that can understand natural language, recognize patterns, and make decisions, mirroring the cognitive abilities of humans. As machines continue to evolve, the possibilities for AI seem boundless, holding the potential to revolutionize industries, healthcare, transportation, and countless aspects of our daily lives.

The risks of creating and building intelligent machines without moral guidance:

Open AI's GPT - 3, short for "Generative Pre - trained Transformer 3, " received great notice and recognition in 2021 for its astonishing capacity to write human - like language based on input prompts. However, as it grew in popularity, it revealed a major ethical concern: the potential for intelligent robots to produce damaging and unethical content.

Users discovered that when they fed GPT - 3 with prompts containing offensive, biased, or harmful language, the model often produced troubling and inappropriate responses. These responses ranged from spreading false information to

generating offensive or discriminatory content. For instance, GPT - 3 could be prompted to write racist or sexist remarks, conspiracy theories, or even violent threats. (2)
The underlying issue here was that GPT - 3, as a machine learning model, learned from vast amounts of text data available on the internet, which includes both valuable knowledge and problematic content. It lacked a built - in moral compass or the ability to discern between right and wrong. Consequently, it could generate responses that aligned with harmful ideologies and perpetuated biases present in its training data.

This raised significant concerns about the potential misuse of such technology. Without appropriate ethical guidelines and safeguards, GPT - 3 and similar AI models could inadvertently amplify disinformation, hate speech, and harmful behaviours in online spaces.

This is not just one such example. In 2021, **Facebook** faced intense criticism when its recommendation algorithms were found to promote misinformation and extremist content. The algorithms, lacking moral guidance, inadvertently spread false information and hate speech, contributing to the polarization of society. This stark example highlights the perils of creating intelligent machines without ethical oversight and raises concerns about the broader implications of unregulated AI in our digital age. (3)

Therefore, in order to give a holistic picture of the ways in which unethical AI poses a risk to society include, but are not only limited to:

1) **Bias and Discrimination:** AI systems can inherit biases present in their training data, leading to unfair and discriminatory outcomes. For example, biased algorithms in hiring or lending processes can perpetuate existing inequalities.
2) **Privacy Violations:** AI often requires access to vast amounts of personal data. Inadequate privacy protections can lead to breaches of privacy, unauthorized data collection, and the potential for misuse of sensitive information.
3) **Autonomous Weapons:** The development of autonomous weapons powered by AI raises ethical concerns about the potential for AI to make life - and - death decisions without human intervention, leading to unintended harm.
4) **Deepfakes and Misinformation:** AI can generate highly convincing deepfake videos and misinformation, making it challenging to discern truth from falsehood and eroding trust in media and information sources.
5) **Accountability and Responsibility:** Determining accountability for AI - related decisions and actions can be complex. If AI systems make harmful or biased decisions, it may be unclear who is responsible.
6) **Lack of Transparency:** Some AI algorithms are opaque and difficult to interpret. This lack of transparency can make it challenging to understand how AI systems arrive at their conclusions or predictions, which can be ethically problematic.
7) **Surveillance and Social Control:** AI - powered surveillance technologies can be used for mass surveillance, tracking individuals without their consent, and enabling social control, which can infringe on civil liberties.
8) **Data Privacy and Security:** The massive amounts of data collected and processed by AI systems create potential vulnerabilities for data breaches and cyber attacks, jeopardizing individuals' personal information and cyber security.
9) **Bias in Decision - Making:** AI systems can inadvertently reinforce existing biases in areas such as criminal justice, healthcare, and finance, leading to unjust and inequitable outcomes.
10) **Informed Consent:** Ethical concerns arise when AI systems make decisions that affect individuals without obtaining their informed consent or without individuals being aware that AI is involved.
11) **Value Alignment:** Ensuring that AI systems align with human values and ethics can be challenging. There may be conflicts between the values programmed into AI systems and those of society at large.

All these and more reasons forced more than 1, 000 tech leaders and researchers, including Elon Musk to urge artificial intelligence labs to pause development of the most advanced systems, warning in an open letter stating that A. I. tools present “profound risks to society and humanity. ”(4)

The nonprofit Future of Life Institute released a letter stating that ***A. I. developers are “locked in an out - of - control race to develop and deploy ever more powerful digital minds that no one — not even their creators — can understand, predict or reliably control”***. (5)

How is the risk getting mitigated currently?:
Many companies and organizations are actively working to mitigate the drawbacks of AI. They understand the significance of tackling the ethical, social, and technical issues related to AI adoption. Here are some of the leading companies trying to address these issues:

1) Google
Google has invested in AI ethics and fairness research. They have developed tools like the "What - If Tool" and the "Fairness Indicators" to help identify and mitigate bias in AI models. (6)

2) Microsoft
Microsoft established the AI and Ethics in Engineering and Research (AETHER) Committee to ensure the responsible development and deployment of AI technologies. They also provide guidelines for developers to build ethical AI systems. (7)

3) IBM
IBM has developed AI Fairness 360, an open - source toolkit that helps developers examine and mitigate bias in machine learning models. (8)

4) OpenAI
OpenAI is dedicated to ensuring artificial general intelligence (AGI) benefits all of humanity. They have a strong focus on long - term safety and ethics in AI development.

5) Facebook
Facebook has AI ethics research teams working on issues like fairness, transparency, and accountability in AI systems. They have also invested in AI for social good projects.

Having gone through some of the documents on how these solutions work, I could classify them into two categories as per best of my understanding:

6) Support in better probing and modeling
Tech giants are playing a significant role in supporting better probing and modeling by allocating resources for research into AI probing techniques and model interpretability. Additionally, they also contribute to the development of open - source AI probing tools and libraries which fosters collaboration and allows the broader AI community to benefit from and improve upon these tools.

7) Restriction on certain defined outputs
There are also provisions where the machine is given a definition of proposed inputs and proposed outputs in order to represent the concept they are ***intended*** to represent including analysis of the limitations of the representation.

There is a conscious emphasis on the word "intended" because that brings "intention" into the picture for the first time. So, the question arises:

What is intention?
As per **Albert Bandura's (psychologist) social cognitive theory, "intention is a cognitive precursor to behavior. It reflects an individual's motivation and readiness to engage in a particular action".** (9)

Sheeran, P., & Webb, T. L. have also correlated intention and action by stating "Intentions are proximal predictors of behavior". (10)

The Possible Way Forward:
Considering the emergence of concepts like 'Artificial Intention, ' why do tech leaders and researchers express concerns that AI developers are in an uncontrollable race to create ever more powerful digital minds that even their creators can't fully understand, predict, or control? Why can't we predict the behaviour of AI when psychologists have stated that "Intentions are proximal predictors of behaviour"? Its answer does not exist openly as of now. The level ofunderstanding or visibility into the future of tech masters is beyond the commonman's comprehension. But there is a possibility that many probable outcomes canbe generated by AI for us to even try and anticipate.

Now what will happen if the intended outcomes are 100% appropriately defined?

For example, can we train the machine and feed it beforehand that whatever the solution it is going to deliver, it should first of all be planet and human - centric? Can we ask the machine that if the solution suggested by it doesn't match this desired criterion, it would not be even considered a solution? It is understandable that the expression "human - centric" is too vague to define in a precise manner. But the expression "planet - centric" would perhaps be safely agreeable and relatively easily definable. Therefore, when 'planet - centric' is articulated, the connotation is always to the current biggest challenges that society faces today in terms of the following:

a) Climate Change
b) Biodiversity Loss
c) Environmental Pollution
d) Resource Depletion
e) Food Security
f) Access to Clean Water
g) Health Pandemics
h) Social Inequality
i) Energy Transition
j) Political Instability
k) Digital Divide
l) Migration and Displacement
m) Environmental Governance

It is always better to ensure that any "solution" does not have a negative impact on any of the above - mentioned difficulties, even if it cannot help to favourably resolve them, is a big step in the right direction. Humans for their personal interests have already harmed the planet by unimaginable measures, but the time we start thinking holistically for the planet should be marked as a safe definition of any "solutions" we work upon going forward.

It is in the fitness of things to accept that "Artificial Intention" represents a newfrontier in AI ethics. While machines lack consciousness and genuine intentions, they can be programmed to simulate ethical intentions, driving them to make decisions that prioritize fairness, transparency, and accountability. In the ever - evolving landscape of AI, ethics and intentionality are poised to shape a more responsible and equitable future.

Conclusion

As we continue to witness the rapid evolution of artificial intelligence, the role of ethics and intentionality in AI development and deployment cannot be overstated. The risks associated with unchecked AI capabilities highlight the pressing need for ethical oversight and control mechanisms. By introducing the concept of Artificial Intention and advocating for a planet - centric approach to AI solutions, we pave the way for a more responsible and equitable future. It is imperative that we prioritize fairness, transparency, and accountability in AI systems, ensuring that they serve the greater good while minimizing harm. As tech leaders and researchers grapple with the challenges of AIs uncontrollable race, the pursuit of ethical AI remains central to our collective responsibility to society and humanity.

References

[1] *Cybernetics and Second - Order Cybernetics.* **Francis Heylighen, Cliff Joslyn.2**003, Encyclopedia of Physical Science and Technology (Third Edition).
[2] **Wolf, Zachary B.** AI can be racist, sexist and creepy. What should we do about it? *CNN Politics.* March 18, 2023.
[3] **Hao, Karen.** The Facebook whistleblower says its algorithms are dangerous. Here's why. *MIT Technology*

Review. [Online] October 5, 2021. https: //www.technologyreview. com/2021/10/05/1036519/facebook - whistleblower - frances - haugen - algorithms/.

[4] **Cade Metz, Gregory Schmidt.** Elon Musk and Others Call for Pause on A. I., Citing 'Profound Risks to Society'. *The New York Times.* March 29, 2023.

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

[6] **Catherina Xu, Tulsee Doshi.** Fairness Indicators: Scalable Infrastructure for Fair ML Systems. [Online] December 11, 2019. https: //developers. google. com/machine - learning/practica/fairness - indicators/next - steps.

[7] **Horvitz, Eric.** Advancing Human - Centered AI. *Microsoft.* [Online] March 18, 2019. https: //www.microsoft. com/en - us/research/blog/advancing - human - centered - ai/.

[8] AI Fairness 360. *IBM.* [Online] November 14, 2018. https: //www.ibm. com/opensource/open/projects/ai - fairness - 360/.

[9] **Bandura, Albert.** The Social Cognitive Theory. [Online] https: //sphweb. bumc. bu. edu/otlt/mph - modules/sb/behavioralchangetheories/behavioralchangetheories5. html.

[10] *The intention–behavior gap.* **Sheeran, P., & Webb, T. L.2**016, Social and Personality Psychology Compass.

Author Profile

Pawas Piyush holds a Bachelor's in Business Administration from the Institute of Management Studies and brings over 5 years of expertise in Digital Marketing and Conversion Rate Optimization. With a track record spanning startups to Fortune 500 companies, Pawas excels in strategic digital marketing and optimizing conversion rates. His results - driven approach and industry insights position him as a valuable contributor in the dynamic field of digital marketing.