MERCATUS CENTER
George Mason University

POLICY BRIEF

On Defining “Killer AI”

Nathan Summers and Dr. Sergio Coronado

July 2023

1. INTRODUCTION

The exponential improvements in AI over the past decade have demonstrated its potential to drastically alter current social paradigms. The excitement surrounding ChatGPT has heightened awareness of AI and is leading more people to be interested in how this emerging technology may benefit society. Such excitement is justified. AI systems have been used to better diagnose medical conditions, more efficiently develop pharmaceutical drugs, and increase productivity across a wide range of domains.[1] With the rise of generative systems, such as ChatGPT and Stable Diffusion, humans now have time to focus on high-level tasks, for which AI systems are currently not well suited. This all serves to increase human productivity to a previously unforeseen extent and could lead to social advances on the scale of the agricultural and industrial revolutions.

These benefits, however, do not come without challenges. As concerns about the possible risks of ChatGPT become more apparent with increasing numbers of users, focusing exclusively on generative systems would be a mistake. It is necessary to consider the broader applications of AI systems. Already, different types of AI systems have proliferated into several aspects of daily life, including the automotive industry, the pharmaceutical industry, and social media. These systems can, by their nature, lead to harm, death, or other adverse consequences, even when not explicitly designed to do so. Thus, it becomes important to consider the issues where AI might become deadly: a Killer AI.

The principal issue at hand is to define *Killer AI* in a way that is descriptive and sufficiently robust. On the surface, this seems simple enough. **A Killer AI is a system employing artificial intelligence techniques that, either by design or by circumstance, directly lead to physical harm or death.** A closer reading, however, elicits two important questions: 1) What is meant by *circumstance*, and, more significantly, 2) What is meant by *directly*?

The views presented in this document do not represent official positions of the Mercatus Center or George Mason University.

A creator of an AI system may not have designed it to harm, but it could still lead to an individual's death. Take, for example, the cases wherein self-driving cars have killed both their passengers and nearby pedestrians.[2] In these situations, a system designed for constructive purposes has, due to a variety of circumstances, led to multiple deaths. Similarly, an individual could deploy this self-driving technology to deliver an improvised explosive device (IED) with the explicit intention of causing harm. This would contradict the original design philosophy of the AI system in question. Consequently, the addition of *circumstance* to a working definition of Killer AI is necessary.

The second of these questions is significantly more complex and presents much of the difficulty in defining and categorizing Killer AIs. How can an AI system directly lead to an individual's death? To answer this, a distinction must be drawn between *virtual AI systems* and *physical AI systems*. *Physical systems* are those which incorporate AI techniques (for tasks such as navigation, vision, etc.) into physical hardware and can thus interact directly with their environment. *Virtual systems* refer both to these internal AI systems as well as AI systems which were not designed to directly interact with the physical world. Regarding lethal autonomous weapon systems and other physical AI systems—even those not designed for explicitly lethal purposes—the answer is obvious. However, in content aggregation algorithms, hiring algorithms, engineering algorithms, and other virtual AI systems, the distinction between direct and indirect harm is less clear. This is because an algorithm cannot itself engage in a physical action. Instead, some other mechanism must exist to inflict harm on humans. While this might seem to preclude a virtual AI system from "directly" harming or killing, this is not necessarily the case.

Legal liability may provide an answer to this question. In cases of liability, proximate cause must be established. *Proximate cause* is a legal term that refers to a cause of an injury or harm that is legally sufficient to determine liability. It can be summed up with the phrase *but for*, i.e., "but for this act, the injury would not have occurred."[3] The nature of this cause will vary depending on the circumstances of the case, such as whether the liability being argued is civil or criminal. Applying this to the issue at hand, an AI system must contribute a sufficiently significant cause to be considered directly responsible.

Already, there have been instances where proprietors of AI systems have faced accusation of being responsible for the death and injury of individuals—and may be held legally liable.[4] However, this elicits questions regarding on whom the liability should fall in the case of civil or criminal nonadherence. Does the responsibility of ensuring the safety of deployed AI systems rest with legislators, the organization deploying the algorithm, the developers, the users, or some combination thereof?

Most notable is the case of Molly Russell, a young girl who took her own life in 2017.[5] The coroner's inquest found that Instagram's and Pinterest's content aggregation algorithms flooded her social media with a romanticized view of self-harm and suicide.[6] While the algorithm itself could not physically kill Molly Russell, it played a role in her death; therefore, the social media sites bore some responsibility.

More hypothetical is the case of a repurposed pharmaceutical-seeking algorithm.[7] In an experiment, a researcher repurposed and tasked an AI system to identify potentially toxic chemical compounds. The algorithm generated 40,000 possible chemical weapons in the span of six hours. Obviously, the algorithm cannot synthesize and deploy chemicals itself, but it supplied the deadly information. In the wrong hands, any one of these 40,000 might be used in a chemical attack. Would the creators be "directly" responsible for the resultant deaths?

These cases provide a compelling argument that virtual AI systems *could* be directly responsible for human harm and death; however, the cases are insufficient to define *how* responsibility for such can be determined. To resolve this question, we propose a rigorous framework to determine the precise point at which a virtual AI system is directly responsible for human harm and death. Such a definition must be sufficiently descriptive and robust for both the academic and general population to adopt it.

We advocate for this standard: **a virtual AI system bears direct responsibility if the number of subsequent reactions to its output required to inflict harm is below a threshold proportional to the potential severity of said harm.** This approach takes inspiration from proximate cause in cases of legal liability. A *subsequent reaction* is a human action taken after receiving the output of the AI system. An AI system that requires only one subsequent reaction in order to inflict harm bears direct responsibility for the harm. As the severity of harm increases, more subsequent reactions are required to clear the AI system of culpability. In the case of the repurposed pharmaceutical algorithm, for example, the threshold would be significantly higher as the potential lethality of a chemical attack elevates. This means that the AI system would be liable for any harm caused by the list of chemical weapons unless it were determined that the user took enough subsequent reactions to clear the AI system of culpability.

This briefing will primarily focus on both defining these subsequent reactions and providing a framework that establishes a threshold for Killer AIs.

2. ALGORITHMIC UNCERTAINTY

Before defining reactions and determining appropriate reaction thresholds, an important caveat to this definition must be discussed. The Molly Russell case demonstrates that AI systems can lead to human harm or death, even when there is not a clear physical component of the system's output, as there is in the chemical weapon example. This is because of how the underlying mechanisms through which AI systems reach their decisions work.

When a person tasks an AI system with reaching a decision, the system assesses available data and determines the "correct" output by comparing current data to its training data. In some situations, the training data and the available data will not line up exactly. This affects output. Even when situational data are exactly equal to training data, other confounding factors exist. These factors might not have been taken into consideration during training, or the system might be incapable of detecting them. Consequently, an AI system's output is never made with 100 percent certainty (although it may come close). Output is an expression of what the AI determines to be the most statistically likely solution given a novel situation.

To explain this point more simply, consider a simple AI system tasked with distinguishing and displaying images of cats and dogs. The system will have two possible output categories—one for cats and one for dogs. It will consider a number of features from its input image when determining which animal to display. The AI programmer plays a crucial role. How and with what data a programmer trains the model will determine which features the system will consider and how those features will factor into the analysis. This setup will impact the system's final determination.

For example, the system might learn to distinguish the animals by focusing on the eyes. (It is important to note, however, that the features selected by AI systems will not necessarily be easily recognizable to humans. The example of eyes and pupil shape used here is for illustrative purposes and should not be taken as indicative of the types of features actual systems will select.) Vertically oriented pupils correlate with cats; round pupils correlate with dogs. However, images of cats taken in darker conditions may display rounder pupils. This demonstrates how variations in a single feature might introduce uncertainty. Combine that variance with hundreds or thousands of features, and it becomes clear how systems—even simple ones—might generate uncertainty and provide incorrect or inappropriate outputs.

While algorithmic uncertainty is apparent in virtual systems, it also exists in physical ones. This is because the physical system relies on underlying virtual AI systems, allowing these virtual systems to interact directly with the physical world without being prompted by a human.

Algorithmic uncertainty has another challenge called the *alignment problem*. Because of the way programmers assign goals, AI systems may determine that the best solution is one that a human would never conclude. This conundrum might come when an AI system seeks to address concerns over climate change. The hypothetical AI system determines that the best solution, instead of limiting carbon emissions, designing smarter cities, increasing the availability of public transport, etc., is to eliminate the root cause of climate change: humans. While this example is extreme, it illustrates the alignment problem.

A simpler example is using AI in self-driving cars. If the programmer prioritizes arrival times, the AI output might be algorithmically incentivized to ignore traffic laws in order to optimize travel time. Inadvertently, this prioritization may place people (pedestrians, the car's own passengers, and occupants of other cars) in danger. Consequently, it is important to incent prioritization of other aspects important to travel, such as pedestrian safety, passenger safety, and lawfulness. Even then, however, the algorithm driving a car must choose which of *these* aspects to prioritize.

AI systems affected by algorithmic uncertainty still fall under the overarching definition for Killer AI: they are systems employing artificial intelligence techniques that, either by design or by circumstance, directly lead to physical harm or death; however, the uncertainty, in a sense, may make physical Killer AI immune from considerations regarding reactions and reaction thresholds. For example, with lethal autonomous weapon systems, harm caused through algorithmic uncertainty does not depend on any external agency or subsequent reaction. This remains an important aspect of Killer AI—one that deserves consideration when assessing the potential of harm from an AI system.

3. DEFINING REACTIONS

The most significant challenge in the proposed approach is the ambiguity surrounding the term *reaction*. In the case of the repurposed pharmaceutical algorithm, there are two clear and necessary reactions that must occur before any harm can be inflicted: the synthesis of the chemical weapon and its subsequent deployment. However, each of these steps can be broken down infinitely. The synthesis of the weapon, for example, has several parts: acquiring component chemicals, acquiring laboratory materials and space, hiring or training chemists, etc. Each of these resultant steps can also be broken down further, instigating a process which may continue ad nauseum.

Therefore, a reaction is a comprehensive combination of sub-steps that stand alone from other actions. In the case of weapon synthesis, combining subsequent actions into a single step would be logical. Synthesizing a chemical weapon cannot, however, be combined with its deployment, because its deployment is necessarily dependent upon its synthesis. Consequently, a single reaction must depend on a preceding action and must be necessarily required for a succeeding action. In other words, each reaction must have a clear prerequisite (which may include the initial output of the AI system) and post-requisite (which may include the resultant human harm or death).

Within a single reaction, the principle is that no temporal dependency can exist between one sub-step and another, as the chemical weapon synthesis example demonstrates. The acquisition of chemical components can happen before, after, or concurrently with the acquisition of laboratory materials and space, so synthesis must all be one step.

Behind the scenes, someone or something—a human, a mechanical or robotic system, or some subsequent AI system—must act as the decision maker. In other words, for a reaction to occur, something (human or otherwise) must make the decision to react and then carry out the reaction in the physical world. Because of the nature of the virtual AI systems currently under investigation, this agent must be a human, but a virtual AI system may have recruited the human agent. (This was the case with GPT–4[8]). In the chemical weapon example, the synthesis of the weapon could occur by a human agent or a mechanized chemical synthesis platform.

The above reasoning can be distilled into the following definition: **a single reaction is a sequentially dependent step that requires an external decision-making agency between the output of an AI system and eventual harm. This step cannot be combined with another reaction.**

Applying this definition to the Molly Russell case, a single subsequent reaction is obvious: she took her own life. The social media content (the output of the AI system) that Russell consumed romanticized self-harm and suicide. Because of the content that was pushed on her by social media content aggregation algorithms, AI contributed to Miss Russell's state of psychological distress and led to her suicide.

The induction of psychological distress is not its own reaction, because Russell did not actively choose to become psychologically distressed. Instead, the content given to her by Killer AI proved distressing enough to induce a mental state outside her control. Consequently, this psychological state cannot be considered to result from external agency.

4. DETERMINING THE REACTION THRESHOLD

Having arrived at a definition for a reaction, determining the threshold of subsequent reactions for an AI system to be considered directly responsible for harm comes next. This threshold is necessary, as an increase in the number of subsequent reactions required to inflict harm permits an increase in confounding variables. A system that might eventually inflict harm but requires an excessive number of subsequent reactions to do so may no longer meet the "directly responsible" threshold. Factors outside of the AI's sphere of influence may play a sufficiently large role, requiring the distribution of responsibility across several actors.

However, it is important to scale this threshold with the potential harm that an AI system could cause. As an AI's potential harm becomes more lethal to more people, it may become necessary to consider more reactions. Accounting for the higher accompanying risk and ensuring AI systems with far-reaching impact remain safe may require flexibility of the "directly responsible" threshold.

When AI systems harm a limited number of people, it is necessary to consider only a few subsequent reactions. However, when AI systems induce psychological distress, various comorbidities may also contribute to decisions to inflict harm. The further removed the eventual harm is from the AI's output, the more profound the impact of the comorbidities will be. Consider this hypothetical case: an emotionally troubled person reads political disinformation and propaganda on their social media feed. They become radicalized and then plan and execute an act of terror. It is fair to say that content aggregation algorithms played a role in their radicalization; however, other preexisting mental health conditions also contributed to their deadly actions. Assigning direct responsibility largely depends on the level of human agency present in the planning and perpetration of the act. Therefore, this hypothetical case is different from the Molly Russell case, if for no other reason than the preexisting mental state of a potential terrorist is likely more complex than that of a suicidal person.

To provide a framework for establishing a reaction threshold, one must determine which outcomes of Killer AIs are least desirable. It may seem obvious, but we argue that death, regardless of the number of casualties, is the least desirable outcome when compared to injury or harm. Furthermore, when death is uninvolved, the well-being of many people prevails over the well-being of one or just a few. Bluntly put, people can recover from nonfatal injuries.

Accordingly, we have developed four categories of potential harm in ascending order of severity:

1. Harm to one or few
2. Harm to many
3. Death to one or few
4. Death to many

Thus, AI systems with potential impacts later in the list (categories three and four) compel deeper consideration than those earlier in the list (categories one and two).

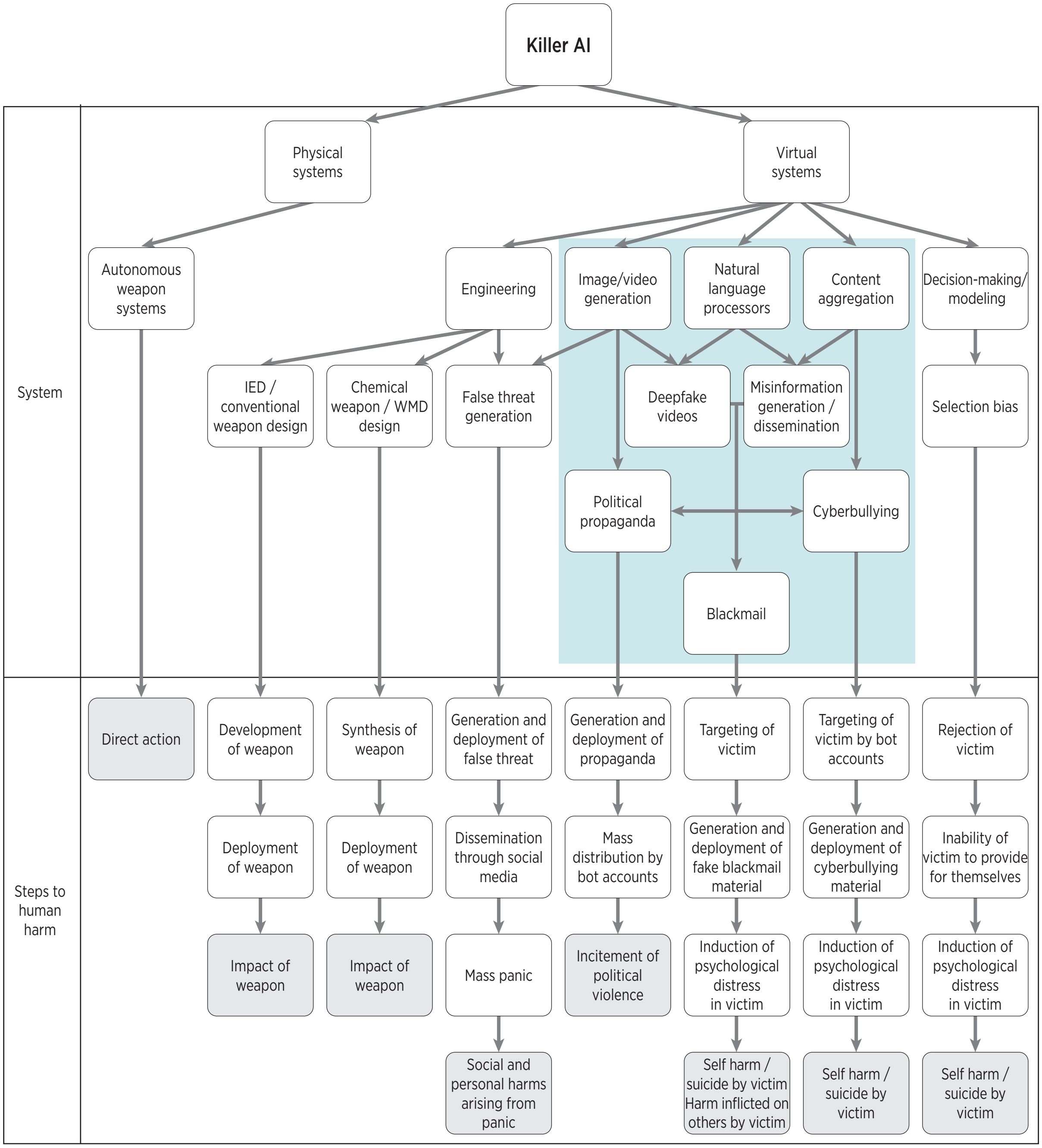
Naturally, the reaction threshold will vary on a case-by-case basis. Consequently, determining which potential harm category an AI system falls into requires an interdisciplinary group's analysis. This group should represent a wide range of vested interest in the deployment of the AI system in question so it is not swayed by a single biased viewpoint. For example, a group with significant financial interest in the deployment and adoption of a specific system may overlook (or depreciate) potential severity. We recommend a group to include developers, academics, and sufficiently educated legislators to provide a well-rounded and objective determination of system severity.

5. SCOPE

Figure 1 illustrates the far-reaching scope of Killer AI systems as well as prerequisite steps required for systems to inflict harm. It should be noted that this figure is an early draft. As such, it should be considered non-exhaustive and subject to future changes.

Outlining this scope is relevant, as AI systems vary significantly in their application, capabilities, and potential for harm. Furthermore, because civil applications of AI may be capable of inflicting harm, even unwittingly, it becomes necessary to establish a scope of the types of AI systems that may cause harm—and identify potentially harmful avenues. This scope, and others like it, may guide policymakers, developers, and other stakeholders in determining appropriate reaction thresholds for different types and applications of AI systems.

Figure 1. Working draft of the scope of Killer AI


Killer AI
System
Physical systems
Virtual systems
Autonomous weapon systems
Engineering
Image/video generation
Natural language processors
Content aggregation
Decision-making/ modeling
IED / conventional weapon design
Chemical weapon / WMD design
False threat generation
Deepfake videos
Misinformation generation / dissemination
Selection bias
Political propaganda
Cyberbullying
Blackmail
Steps to human harm
Direct action
Development of weapon
Synthesis of weapon
Generation and deployment of false threat
Generation and deployment of propaganda
Targeting of victim
Targeting of victim by bot accounts
Rejection of victim
Deployment of weapon
Deployment of weapon
Dissemination through social media
Mass distribution by bot accounts
Generation and deployment of fake blackmail material
Generation and deployment of cyberbullying material
Inability of victim to provide for themselves
Impact of weapon
Impact of weapon
Mass panic
Incitement of political violence
Induction of psychological distress in victim
Induction of psychological distress in victim
Induction of psychological distress in victim
Social and personal harms arising from panic
Self harm / suicide by victim Harm inflicted on others by victim
Self harm / suicide by victim
Self harm / suicide by victim

The blue section of Figure 1 highlights an area of AI development that is rapidly changing and, thus, requires further monitoring. Specifically, this area represents the impacts of generative AI technologies, such as ChatGPT, Stable Diffusion, Midjourney, etc. The intersection of these types of systems—in combination with the prevalence of social media networks—form the foundation for a growing number of potential Killer AIs. These include, among others, concerns about the generation and proliferation of disinformation, an increase in the spread of propaganda, and an increased potential for targeted attacks, such as blackmail, phishing, or cyberbullying. Because generative AI systems are relatively recent developments and their adoption by general society is still in its infancy, the opportunity for risk assessment is also in an infantile stage.

This last point raises a pertinent question regarding whether generative systems can themselves be Killer AIs. We believe they can. In examining the repurposed pharmaceutical system case, research revealed these systems require initial human prompting to begin demonstrating adverse behavior. While some systems, such as ChatGPT, have guardrails built in to prevent users from providing inappropriate prompts, users can circumvent them through creative prompting. Bypassing these guardrails opens the door to taking the media generated by these systems, combining it with output from other systems, and proliferating harmful content over the internet. Currently, these systems require a human agent to feed one system's output into another, but this appears to be rapidly changing. With the introduction of GPT-4 and the ability for users to create plug-ins for ChatGPT, the days of needing human involvement in the process may be numbered.

An early example of this is already apparent with AutoGPT.[9] When users provide Auto-GPT with a goal, the system assigns itself objectives in order to complete the goal. It can then make alterations in light of new information, and then autonomously execute those objectives. This differs from the current configuration for ChatGPT and GPT-4, which require manual commands for each task.[10] Although this particular use of generative AI is still new, it demonstrates that these types of systems are technologically feasible and on the horizon. Thus, generative systems require the same level of scrutiny as other Killer AIs.

6. POTENTIAL LIMITATIONS OF THE CURRENT DEFINITION

It seems appropriate to address limitations and challenges to this Killer AI argument. Some might perceive a logical gap of combining a human's psychological state and the AI system's output into one reaction. Shouldn't they be considered as separate reactions? After all, a psychological state requires an external actor. The definition we argue for, however, includes the requirement of an external *agency*, not necessarily an external *actor*. To say it clearly, a psychological state might not be the result of a conscious decision; it might be induced against one's will.

Furthermore, some might argue the proposed ranking of the categories for potential harm. They might say that an increase in the number of potential casualties caused by an AI system is more significant than the increase in lethality of a system—in other words, that the categories of "harm to many" and "death to one or few" should be reversed within the reaction threshold framework. After all, a widespread nonfatal injury may bring greater financial, societal, and psychological harm to many people than the death of just one person. Nevertheless (and important to the current ranking), injuries often heal, but the loss of even one life is irrecoverable.

A third challenge could focus on the ambiguity of the reaction thresholds framework. One might ask, “Where does ‘few’ stop and ‘many’ begin?” “What category does a system fall into if many people are injured but a few die?” While these are pertinent questions, a fundamental difference exists between a system that can result in death and one that can result in only injury. It is important to conceptually separate these systems when assessing culpability. Asking these types of questions forces developers to consider the nuance implied in the harm threshold framework. In this sense, the framework might benefit from this ambiguity, as it would permit the proposed interdisciplinary committee to approach these issues from a variety of different (even contradictory) viewpoints.

7. FURTHER DISCUSSION

An important note when considering the lethality of any AI system is its reliance on artificial intelligence techniques carried out in purely virtual systems, regardless of whether the system itself results in any physical manifestation. Even in a physical AI system, the underlying decision-making is carried out by virtual AI systems. Accordingly, the considerations made for physical systems must include all those that would be included for virtual systems, with the addition of those that arise from their ability to interact with the physical world without the need for external agency. With self-driving cars, for example, does the car possess the capability to autonomously carry out the actions the algorithms determine: acceleration, breaking, changing lanes, etc.?

Moreover, some systems require initial human input to begin displaying tendencies characteristic of Killer AIs. Content aggregation algorithms are an example of this. Initially, they provide users with a generic collection of content. Over time, by observing which content users interact with and spend the most time on, the algorithm adjusts the curation of content to better fit the specific user’s profile. However, this content does not necessarily need to consciously appeal to the user. Instead, the algorithm might figure out that people interact more with and spend more time on a service whose content makes them angry. In this case, people might be shown more inflammatory content, which has been found to increase how much time they spend using the service but has an adverse impact on their mental state.[11] This demonstrates how an initially benign algorithm might begin displaying dangerous behavior.

Similarly, in the example of the chemical weapons algorithm, the researchers intentionally modified a benign drug-seeking algorithm for adverse purposes. While it is likely that the original developers did not intend for their system to be used in such a way, it remained technically feasible. This demonstrates the importance of considering the widespread impacts of an emerging technology when determining its potential lethality, even when such uses do not align with the system’s initial design philosophy.

These last points raise an important concern given both the newness and potential danger of AI. At what point do the actions of the user become the determining factor in whether a system is a Killer AI? To what extent is the underlying model responsible for potential misuses? This is a pertinent question in the case of generative systems, the drug-seeking algorithm, and even social media content aggregation algorithms. With large language models (LLMs), such as the GPT models, this is particularly salient; third parties can fine-tune them to suit niche applications. If one of these successive models can be used for destructive purposes, should the underlying model be held responsible? The subsequent application developer? The end user?

Another concern relates to the use of reinforcement learning techniques that might change an AI's behavior over time and develop emergent lethal tendencies. This further exacerbates the argument regarding initially benign systems that develop harmful behaviors. Because reinforcement learning techniques often operate without significant human oversight, a change from benign to harmful could occur without notice. A system could engage in harmful or lethal action before developers know to respond.

This harmful behavior might continue in a dangerous direction by biased training data. The AI system's actions would likely reinforce the biases and further propagate them through society. Given the examples of Molly Russell and toxic chemical compounds, the reality of a higher potential for harm elevates the need for caution when permitting the use of reinforcement learning techniques within these systems.

Finally, the importance of algorithmic uncertainty impacting all AI systems compels inspection when considering Killer AIs. All AI systems, physical or virtual, depend on decisions made by underlying virtual systems. Erroneous outputs or misclassifications made by these systems may unintentionally cause harm. Developers, legislators, and users must take greater care to ensure that the systems being developed, regulated, and used are ethically beneficial. The shortcomings inherent in this emerging technology calls for a high standard.

8. CONCLUSIONS

We have put forward the following definition of a Killer AI: **a Killer AI is a system employing artificial intelligence techniques that, either by design or by circumstance, directly lead to physical harm or death.** This is the first known attempt to define this dangerous component of the AI technology, which is becoming increasingly prevalent. We invite further academic research and discussion, more comprehensive legislation, more appropriate regulation, and more nuanced ethical discussions.

From this initial definition, an important distinction has been drawn between physical AI systems and virtual AI systems in order to specify how the latter can, in fact, be directly responsible for harm and death. This determination is based on the concept that **a virtual AI system bears direct responsibility if the number of subsequent reactions to its output required to inflict harm is below a threshold proportional to the potential severity of said harm.** On this basis, AI systems that were previously incapable of being considered directly responsible for harm can now be assessed more rigorously.

However, to assign direct responsibility to AI systems, a definition for a reaction is also required. **A single reaction is a sequentially dependent step that requires an external decision-making agency between the output of an AI system and eventual harm. This step cannot be combined with another reaction.** Using these definitions enables a better classification and understanding of not only *how* AI systems may lead to harm but insight into *why* they might cause harm. It also provides clarification regarding the role of confounding factors in inflicting eventual harm.

Finally, we have proposed a framework through which to assess and classify the harm that may be caused by potential Killer AIs. This framework posits four categories of increasing harm, organized around the premise of preserving human life.

In light of these contributions, this briefing aims to bring attention to an area of significant concern within the field of AI research. Furthermore, it seeks to stimulate conversation on the responsibility of AI systems and their developers, legislators, academics, and users with regards to the potentially lethal dangers of AI. Moreover, we hope that these contributions encourage further study on Killer AI. Future research in this field could focus on the potential inevitability of these systems, ways these systems could reliably be made safe, and other unforeseen solutions to the concerns we have raised.

ACKNOWLEDGMENTS

The authors of this paper would like to acknowledge the Mercatus Center at George Mason University for their funding and review of the current research, as well as Laura Plein (Luxembourg Tech School) and Mika Beckerich (Luxembourg Tech School) for their feedback and contributions throughout.

ABOUT THE AUTHORS

Nathan Summers is a recent master's graduate with a profound interest in the ethical impacts of emerging technologies. He is currently an AI research analyst with the Luxembourg Tech School studying the potential lethality of AI systems, considering their increasing prevalence in daily life. His specific focus is on proposing ethical solutions which aim to mitigate the harm that can be caused by AI systems, both in civil and military contexts.He received his MSc from the University College Dublin (UCD) and his BSc from the University of California, Los Angeles (UCLA).

Dr. Sergio Coronado is a recognized international expert in Information and Communication Technologies (ICT) with more than 35 years of experience. His interests lie in ICT strategy, digital governance and leadership, cybersecurity, AI, quantum technologies, and education. Parallel to his career of more than 20 years in NATO, Sergio founded, in his spare time, the non-profit Luxembourg Tech School (LTS). LTS was founded with the goal of inspiring students ages 12–19 to learn and apply technology in a real-life context. He is also an Assistant-Professor (associé) at the University of Luxembourg, where he teaches Advanced Project Management. Sergio received a PhD in Industrial Engineering, MSc in Software Engineering, MA in Clinical and Educational Psychology, and BSc in Electrical Engineering.

Both authors declare that this work was carried out under the affiliation of Luxembourg Tech School a.s.b.l.

NOTES

[1] Will Douglas Heaven, "AI Is Dreaming Up Drugs That No One Has Ever Seen. Now We've Got To See If They Work," *MIT Technology Review*, February 15, 2023; Don Nguyen, "How AI Can Help Diagnose Rare Diseases," Harvard Medical School, October 18, 2022; Debleena Paul et al., "Artificial Intelligence in Drug Discovery And Development," *Drug Discovery Today* 26, no. 1: January 2021, 80–93; Sara Reardon, "AI Chatbots Can Diagnose Medical Conditions at Home. How Good Are They?," *Scientific American*, March 31, 2023.

[2] Daisuke Wakabayashi, "Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam," *New York Times*, March 19, 2018.

[3] "Proximate Cause," Legal Information Institute, Cornell Law School, accessed June 23, 2023, https://www.law.cornell.edu/wex/proximate_cause.

[4] Ava Asher-Schapiro, "Lawsuits Pile up As U.S. Parents Take on Social Media Giants," Thomson Reuters Foundation, February 8, 2023; Argus Crawford and Bethany Bell, "Molly Russell Inquest: Father Makes Social Media Plea," *BBC News*, September 30, 2022; Samantha Murphy Kelly, "Their Teenage Children Died by Suicide. Now These Families Want to Hold Social Media Companies Accountable," *CNN Business*, last modified April 19, 2022; Social Media Victims Law Center, "Press Releases," accessed June 23, 2023, https://socialmediavictims.org/press-releases/.

[5] Crawford and Bell, "Molly Russell Inquest."

[6] North London Coroner's Service, *Regulation 28 Report to Prevent Future Deaths*, October 13, 2022, https://www.judiciary.uk/wp-content/uploads/2022/10/Molly-Russell-Prevention-of-future-deaths-report-2022-0315_Published.pdf.

[7] Rebecca Sohn, "AI Drug Discovery Systems Might Be Repurposed to Make Chemical Weapons, Researchers Warn," *Scientific American*, April 21, 2022.

[8] Kevin Roose, "GPT-4 Is Exciting and Scary," *New York Times*, March 15, 2023.

[9] T. B. Richards, "Auto-GPT: An Autonomous GPT-4 Experiment," GitHub, accessed July 7, 2023, https://github.com/Significant-Gravitas/Auto-GPT.

[10] Kevin Jiang, "What's Auto-GPT? New, Autonomous 'AI Agents' Can Act on Their Own, Rewrite Their Own Code," *Toronto Star*, April 14, 2023.

[11] William J. Brady et al., "How social learning amplifies moral outrage expression in online social networks," *Science Advances* 7, no. 33 (2021); Ari Shapiro, Michael Levitt, and Christopher Intagliata, "Social Media Can Inflame Your Emotions — and It's a Byproduct of Its Design," *NPR*, September 6, 2022, sec. Author Interviews.