How do AI Systems Protect Children from Engagement Harm?

Carla Jaffal

How do AI Systems Protect Children from Engagement Harm?

Artificial intelligence has managed to change our digital world, especially changing how people interact with technology. This rise in AI didn't just transform our world, but also our children's. It ranges from the AI-oriented algorithms on their daily apps to the smart toys they use and even the creative games they spend hours on, AI has become integrated in children's everyday lives. These systems are made specifically to be personalized and even adaptive to use users' everyday behaviors to predict their preferences and cause addictive interactions. Even if this kind of interaction provides educational benefits and even help advance creativity, it still draws out a big concern known to be as "engagement harm". Engagement harm refers to the negative consequences of technology and its systems. This kind of engagement focuses on two main aspects: attention and interaction, which for children, this can become even more integrated upon excessive screen time, exposure to inappropriate content, and even loss in child creativity. These effects aren't necessarily caused purposely; however, it can be due to how AI systems are engineered to prioritize engagement. Children are

vulnerable to these algorithms. Their interaction with AI can shape their way of thinking and even their creativity. Here is where the importance of balance between AI's algorithms and protecting children's well-being comes in. My exploration of this topic is inspired by the severity of making sure that AI systems serve as protective means rather than sources of harm for children. At first, I was optimistic about the ability of AI to protect children. However, the more I delved into different sources that highlighted both aspects of AI, I realized how complex it really was. This essay will walk you through my journey through the different viewpoints through different sources, revealing my understanding of the question: How do AI Systems Protect Children from Engagement Harm?

The first source I read was the UNESCO article "How should Children's Rights Be Integrated into AI Governance?" (UNESCO, December 10, 2023). The article focused on how AI systems shape children's experiences by their specific content that personalize their learning and develop their online interactions. Although AI provides various educational benefits and enhances child creativity, it also raises concern about how the data is used ethically and the consequences of engagement-oriented

algorithms. As I was reading, I discovered that AI uses tools that are programmed to find harmful content, remove inappropriate information, and find any risky behaviors. For example, AI algorithms in most social media apps can detect and remove any child exploitation content to protect children from being exposed to any harm. Today, governments, organizations, and researchers are paying more attention on how to protect children from harmful AI engagement. UNESCO talks about how important it is to prioritize children's rights into AI policies, and encourages the idea of providing their best methods to ensure kids' online safety (UNESCO, 2023). However, it might be difficult to provide a good balance. For example, on the positive side, AI can provide various learning opportunities for kids that will help in their development. Still, we cannot deny how it can easily cause addictive behavior. It was even proved by the Conference of the new Council of Europe how AI's effect on the development of children is still considered to be "scattered" as it is not well balanced (Rome, 7⁄8 April 2022). This constant unstability made me realize that we need more rules in AI and more involvement from social media platforms to keep the youth safe. Here comes one of the most important roles of AI in child protection where it has the ability to analyze big amounts of data quite fast. This allows it to identify repetitive harmful

behaviors to reduce the threats (signs of bullying or grooming by predators). These systems are now used on YouTube and TikTok to fastly remove harmful comments and restrict content. This ensures that children are less likely to feel any sort of uncomfort while enjoying their online spaces.

As I continued further in my exploration into how AI systems work in protecting children against harm, I encountered an important case study "AI-The Social Disruption" by Vaithianathan et al. (2021). This article opened my eyes to the Allegheny Family Screening Tool known as (AFST), which is a machine learning tool that is used to guide social workers in targeting children maltreatment hotline calls. Knowing that the tool is designed to identify the e cases where children are at high risk of mistreatment to flag those cases was very promising to me. The thought of using AI to improve such sensitive decisions was very impressive. However, as I delved deeper into the article, I started to wonder if using these tools themselves might cause the engagement harm, or even accidentally make the problems they're trying to fix even worse.

AFST is known for its incredible role in incorporating human design elements, such as leadership, community, and most importantly transparency (Vaithianathan et al., 2021). These elements helped it gain social trust, which is crucial when it comes to tools that rely heavily on algorithms. As an example, when experts from Illinois and New Zealand tried to use similar tools, it led to huge public attack because of fears of privacy invasion and bias. The lack of transparency made it difficult for the public to understand or support the systems. Not like ASFT that involved the community and gave the local agency control over data and algorithms. However, the article itself didn't ignore the risks from this tool. A criticism that stuck with me was that of Virginia Eubanks who argued that such tools could be considered to be "poverty profiling," (targeting low-income families). This made me consider whether gathering big amounts of personal information to identify risks could also be used to harm people by increasing the probability that children from specific backgrounds will be affected by online threats. Even though ASFT showed real active proof of reducing bullying and reducing the use of racial languages, it still sparked in me the question whether fairness and efficiency do come hand in hand. I was very inspired by the tool's well-made design, yet I was also concerned about its potential misuse. Could

the same features that make these tools effective (speed, accuracy, data collection) also make them dangerous to some groups of people?

This source made me expand my perspective beyond the question of whether AI can prevent harm from children, and toward a more complex point of view. It seems the way AI is made in social systems matters just as much as its technical performance. An intention of protecting does not guarantee actual protective outcomes. AI systems like the AFST offer a possible path forward in reducing harm, even harms from things like stress or neglect, but they also challenge us to think about ethical design, transparency, and the consequences of data. As with my earlier reading from UNESCO 2023, which talked a lot about the active role of the government, this article focused on how context and community involvement can indeed influence the success or failure of child protection in AI.

The third source I read was the article “Toward Children-Centric AI”, La Fors (2022) talks about how AI can help not just to protect children from harm, but also to promote their growth and development. The article stresses on how AI interactions with children can be looked upon through the perspective of developmental growth by

distinguishing between actions that promote discrimination and negative actions and positive biases that promote inclusivity and acceptance. It made me change my perspective on protection itself. I initially started with the topic assuming that bias in AI was entirely harmful. However, La Fors introduced an interesting perspective, that is children, particularly between ages 7 and 11, are capable of perceiving bias in their interactions and can even use it as a learning opportunity when guided correctly.

This made me think differently about how engagement harm might arise. As it could form from exposure to negative content or even manipulation through some algorithms in the systems children use. For example, if an algorithm only shows content of one culture or gender stereotype, children may think of these limits as social norms. La Fors suggests that AI systems could be used to work on strengthening positive biases to teach diversity and help reduce the harm caused by stereotypes. That was an entirely new idea for me. This reminded me about the earlier claim in my essay when we considered treating AI as a protective mechanism, but this time with an ethical approach.

La Fros talks about how AI treats children as passive users rather than active participants. This limits the ways children's ability to learn and grow through active

interaction. If AI neutrals out all biases, this might also not allow children to experience and learn from diversity. He even suggested the idea of “co-creational” spaces to allow children to design the AI system and algorithm they interact with. I found this idea very interesting as it focuses on the idea of participatory design and child agency.

This perspective made me question: Can we expect AI to play an educational role in children’s lives? While the previous article explored how community engagement does help to form trustworthy systems, La Fors focuses on how children themselves might shape these systems. Both perspectives seem to have the same approach: It is not a technical fix, Instead, protection must also aim to include educating children to form healthy relationships with technology. This gave me a new perspective that it is important to consider AI as a partner rather than a safeguard.

For my last source, I referred to Wang et al.’s (2024) article "Challenges and Opportunities in Translating Ethical AI Principles into Practice for Children." This source focused less on AI-tools and more on the guidelines of AI design.

One of the main points the authors talk about is that many ethical guidelines treat children as vulnerable categories, ignoring their diversity and potential evolving. This one-size-fits-all approach may not be developmentally appropriate. As an example, the same privacy guidelines over some topic might not serve a 5-year-old and a 15-year-old equally well. It made me wonder whether AI protections are truly as thoughtful as they need to be. Are they just hiding content randomly to reduce risks broadly, or can they adapt to how children's needs do evolve with time? The article focuses on the idea of designing systems that don't only avoid harm, but also acknowledge children's developmental rights and participation.

Wang et al. discussed just how limited role parents play in AI governance. Even though parents do play a huge role in limiting their children's interaction with screens, it is not deniable that nowadays, children often have far more understanding of the digital environment than their

parents that are assumed to be "digital guides". Here is where I wondered: how can we design AI systems that do protect children but without limiting their potential? Here is where La Fors' co-creational model came to my mind, which aims for children to be part of the design process. I hadn't thought about how easily we might

mistake a system's technical performance for its moral effectiveness. For example, a system might be very efficient in detecting inappropriate images, however, it might still fail to allow children's needs for autonomy or identify information.

This article opened my eyes to the idea of how AI doesn't just need to filter out harmful content or collect only necessary data, but it must also aim to understand how children engage, and how to shape their digital understanding. It did connect all the previous points collected from the previous sources by shwoing how design and the user experience are all crucial factors that go hand in hand when it comes to protecting children from engagement harm, and that real protection requires child-aimed thinking in AI systems. It also reminded me that the absence of harm doesn't necessarily mean the presence of wellbeing.

As I went through this journey of reading multiple perspectives of the idea of AI and child protection, I didn't end up with one final answer, rather, I was given a multiple layered understanding of how complex and important this issue was. At first, I approached the question thinking that I will end up learning about some technical issues that can be fixed, things like filters, better detection, and elevated privacy

policies. However, now I see it as a shared responsibility as it is about cooperation of multiple things like design, transparency, inclusivity, ethics, and even imagination. From UNESCO I learned how important it is for the government to have an active role in ensuring that children's rights are respected in AI. From the Allegheny Family Screening Tool, I learned that transparency and community participation are essential if we want people to trust and understand these systems. I learned from La Fors that, when used properly, bias can be a strong sign of inclusivity rather than a weakness. Additionally, I learned from Wang et al. how important it is to move away from "one size fits all" label and toward more flexible systems that consider various developmental stages and life experiences.

Reflecting on the different points of views I had explored; I came to realize that the question of how AI systems and child protection policies handle engagement harm is far more complex than I had initially anticipated. Rather than classifying AI as either harmful or helpful, the sources have shown that it has a way bigger perspective that depends on several aspects like design, ethical decisions, and the roles of parents and children. From the success of the Allegheny Family Screening Tool to the conceptual

models of child-centered AI growth, each resource presented different ways of protection.

What I have learned is that protection is not a one-size-fits-all concept. AI systems must be designed with policies that prioritize all safety, growth, and even participation. Context awareness and human design that put safety first along with development are essential for AI systems..

I'm still struggling with unanswered questions. Can we really rely on AI to make moral choices regarding the lives of children? How can we really involve kids in influencing the technologies that impact them? This journey hasn’t led me to a clear answer, but rather to a bigger understanding of the idea of AI and child protection. This, I now realize, is what makes exploratory thinking both challenging and necessary.

References

Dalton, E., Chouldechova, A., Putnam-Hornstein, E., Vaithianathan, R., & Benavides-Prado, Using a machine learning tool to support high-stakes decisions in child protection. AI The Social Disruption. https://doi.org/10.1002/j.2371-9621.2021.tb00011.x

La Fors, K. (2022). Toward children-centric AI: A case for a growth model in children-AI interactions. AI & Society, 39(1303–1315). https://doi.org/10.1007/s00146-022-01579-9

UNESCO. (n.d.). How should children's rights be integrated into AI governance? https://www.unesco.org/en/articles/how-should-childrens-rights-be-integrated-ai-governance

Wang, G., Zhao, J., Van Kleek, M., & Shadbolt, N. (2024). Challenges and opportunities in translating ethical AI principles into practice for children. Nature Machine Intelligence,6(4),301–308. https://doi.org/10.1038/s42256-024-00805-x