Can AI Protect Children Online?
Professor Andy Phippen
21 September 2023

In an area such as online harm, particularly when we are considering the rhetoric around their prevention and prohibition, things change very quickly.

While I am confident that between the writing of this and the lecture on the 21st, there will not be a new technology that will indeed prevent any form of harm online, there are always new political views and demands on technology to "do more" in this field, especially with the Online Safety Bill[1] so close to royal assent.

In this talk, I will explore the question: Can AI be used to tackle issues related to online harms?

Clearly, there is much interest in the present time around artificial intelligence, and there is also much media and political interest in online harms and how we might stop them. In exploring current political rhetoric, as I will do when considering some online harms threats and their countermeasures, there are many occurrences where policymakers will make claims about the capabilities of technology, and I have seen many politicians state technology providers could address harms "if they wanted to", and how "clever things" can be done with code.

There has been a prevailing view in this policy area over many years that because online technology underlies the communications and facilitation of online harms, the same providers that have developed these platforms should be able to use code to stop harm. And with artificial intelligence making great, and very visible, strides in its capabilities, it is not unusual to hear "Surely artificial intelligence could solve these issues".

In this talk, I am not proposing a deep dive into the operation of AI systems. What I will do is briefly explore the functioning of modern AI systems or, more specifically, the machine learning approaches that underpin most of the emergent exciting applications of the technology and then present a few use cases around online harms that present challenges to these approaches. Specifically, I would like to explore how problematic the binary rhetoric around this policy area is, and how this potentially detracts from the goal of everyone who works in this area, which is it ensure the online world is accessible to all and risk-free as it is possible to be.

But first, a little about me by way of introduction. Unusually for an academic exploring social and ethical issues around online harm, I have a technical background and hold a computer science degree and PhD. I started my career in an AI research lab in the 1990s. And while I do not code anywhere near as much as I used to, I can still make use of these foundational skills when looking at large data sets. I also remain a Fellow of the British Computer Society and work with them to bring evidence-based arguments to technology policy discussions.

However, over the last twenty years, a lot of my time has been spent speaking to young people about how online technologies affect their lives and contrast those discussions with policy and legislative proposals. I also spend a lot of time working with other actors in the area, such as policymakers, NGOs, educators, or parents to understand their perspectives and wishes. Perhaps importantly, perhaps not, I am also a parent.

[1] https://bills.parliament.uk/bills/3137 [Accessed September 2023]

I raise this because a lot of the work I do presently, motivated by discussions with young people, is to better understand the views and biases adults have when trying to keep young people “safe” online and where these biases come from. As a parent, I see a great deal of media coverage that, if I did not work in the area, particularly talking to young people, I would find extremely alarming.

At the present time, it is difficult to not get excited and/or scared about the capabilities of AI given the highly visible and impressive systems that are emerging. While every student in the country is wondering about the uses of ChatGPT[2], every university in the country is trying to figure out ways of detecting assignments produced from these *Large Language Models* (Wei et al., 2022) and prevent them from being used fraudulently. We are undoubtedly in the midst of an *AI Summer*, of which, if we look historically (Kautz, 2022) there have been three. An AI Summer occurs when developments in research and development transfer into real changes in capability and function. And this AI Summer, few would argue, is the most impactful yet.

However, if we are to learn from history, we might also expect an AI Winter to come at some time in the future. Historically, an AI Winter occurs when the venture capital has been spent, the expectations have not been met and the conclusions drawn are that it is not as good as claimed. Which is, perhaps, unfair. As I have said, with such visible examples it is easy to get excited about the application of AI to everything, particularly if we don’t attempt to understand the underlying technologies and their capabilities.

What is emerging from this AI Summer is what you can achieve with these systems with extremely large processing platforms and bigger data sets than have been used before. Falling out of this processing power are increasingly impressive research and development findings around the models that are used to further refine the models and when problems are addressed in a particular domain, the capabilities accelerate quickly. So, we should anticipate even more impressive Large Language Model systems to emerge in the next couple of years, and content creation systems that produce even more convincing images, video, and audio.

Outside of the public-facing application, there are also extremely impressive developments in what we might refer to as narrow-domain systems, where data is plentiful and consistent in nature. If we consider, for example, the developments in image recognition and tumour diagnosis[3], we have a very specific need and large amounts of data we can use to “train” the recognition system. We also see impressive improvements in the AI of gaming systems, and while automated vehicles are still probably a long way from general use, their application in closed systems continues to develop.

However, there have been less impressive gains in areas that are less easily bound, such as systems in social contexts or those that are global in nature. I will return to the challenges of policing language on social media later, and there have been high-profile examples of the challenges, both functional and ethical, of facial recognition systems[4].

To bring things closer to home, one other part of my day job, and also one of the motivations for this talk, is working with organizations who are within the online harm’s ecosystem, and try to understand how different actors interact, hopefully, for our common goal. Recently, I was working with an NGO that had commissioned a software development organization to help them with a technical intervention to provide support for clients with whom they work around issues related to online harms. “It’s very exciting,” I was told, “They’re using AI”.

However, when pressed on how the AI techniques might be used, the data that might be collected in order to train the AI solution, and the nature of the problem they were trying to solve, given every case has unique and specific elements and is not easily replicable or repeated, there was less clarity. This is certainly not the only time I have heard this and, sadly, the outcomes are rarely the system the procurer envisions or one that

[2] https://chat.openai.com/ [Accessed September 2023]
[3] https://www.cancer.gov/news-events/cancer-currents-blog/2022/artificial-intelligence-cancer-imaging [Accessed September 2023]
[4] The use of facial recognition at Kings Cross Station in London has raised issues of both efficacy and privacy: https://www.theguardian.com/technology/2019/oct/04/facial-recognition-row-police-gave-kings-cross-owner-images-seven-people [Accessed September 2023]

the supplier can deliver. While there is rarely a single reason why these things go wrong, a lot centres on classic software engineering challenges around managing requirements and expectations.

As I have already said, the intention of this talk is not to provide a detailed exploration of the nature of machine learning systems and the leading edge of the research world in this area. However, it is worthwhile to provide a very basic model of how a machine-learning approach happens. Put very simply, a machine learning algorithm "learns" by being fed data, then asked to process that data and come up with original responses based upon a specific need. So, for the above-quoted work around tumour diagnosis, at a basic level, the system is "trained" by being shown many images of tumours, and also images of healthy organs, and is then shown new images and asked to diagnose whether the new image is of a tumour or health organ. This is a very simple high-level explanation, it's actually a lot more complex than that but this is a simple model for illustrative purposes that we can use for the other examples in this talk. In the case of diagnosis systems, given the faster processing available, and larger corpora of training data, as well as improvements in the algorithms themselves, results in this field continue to be impressive.

However, there are also high-profile cases where things do go wrong with these systems, which are sometimes presented in a scenario like "when software goes rogue".

During the COVID lockdowns where school children could not sit examinations, end-of-school assessments were instead carried out by schools and fed into a "normalizing" central system[5]. When it was discovered that children from deprived areas were far less likely to receive top grades, there was significant press and public outrage which the then Prime Minister blamed on a "mutant algorithm". A year earlier, a system used within the UK passport agency[6] that was used to process applications and validate the uploading of photographs became a press story for being far more likely to reject photographs from people of colour. The media rhetoric implied that this algorithm was racist in some way.

The issues related to "bad" algorithms have been well explored for a number of years now (Mehrabi et al., 2022) and generally do not relate to an algorithm achieving independent thought and breaking free from its code base, instead being a problem with either the parameters it had been given (in the case of the qualifications software) or the data it had been trained with (in the case of the passport system). When the training data for an image recognition system has more white faces than people of colour, it is more likely to not recognise people of colour as accurately. The biases, one might observe, lie in those who implement the systems rather than the algorithms themselves (O'Neill, 2016).

As I have stated earlier, while narrow domain systems are achieving great things given larger data sets and faster processing, and improvements to the algorithms used, there are still challenges in complex social systems that are typically messy, comprise millions of users, and are very difficult to place system boundaries around.

We might consider the huge global communication systems that underpin online harms to fall into the more complex category. If I reflect upon the "safety" of online platforms and how it has evolved since my early work with young people at the turn of the century, things are far better than they were. However, do these scenarios make effective use cases for machine learning systems?

As I have already said, one of the challenges in this area is the high-profile nature of cases and the need for policymakers to "do something" as a result, regardless of their understanding of the technical or social complexities of these situations. In exploring the challenges around online harms and automating protection, I will draw upon three use cases each underpinned by a piece of political rhetoric. The intention here is not to mock those making these statements, and attribution will not be made. They are more to use as the means to explore why technical interventions might not be as straightforward as they first appear.

[5] https://www.bbc.co.uk/news/education-53923279
[6] https://www.bbc.co.uk/news/technology-54349538

In the first case, which was a comment during the Prime Minister's Questions following the racial abuse of footballers after the European championship in 2021[7]

> *Last night I met representatives of Facebook, Twitter, TikTok, Snapchat and Instagram and I made it absolutely clear to them that we will legislate to address this problem in the online harms bill. Unless they get hate and racism off their platforms, they will face fines amounting to 10% of their global revenues. We all know they have the technology to do it.*

The nature of a lot of similar rhetoric is the view that there is technology to achieve goals, and providers could do it "if they want to". However, a more detailed exploration results in perhaps a more complex picture than the one offered above. While systems operating on a global scale are very good at symbolic matching – i.e. identifying racist keywords and taking down associated posts, then blocking accounts– they are far more challenged by abuse that might not contain any easily detected keywords. While sentiment analysis (Zang, Wang, and Liu, 2018) has developed over the years, it is still not perfect, particularly when it is applied to a global system that would need to process abuse in many different languages. If machine learning techniques were to be used, the training data that needs to be developed would be challenging if we are to expect it to recognise nuance, sarcasm, satire or simple "banter" between friends. As a result of political pressure and the need to "do more", platforms do increasingly try to implement more aggressive interventions, however, the challenge of false positives is something that will not go away. I am sure I am not the only person reading this to have been presented with the following after a post on Facebook:

> *"Your Comment May Go Against Our Community Standards.*
>
> *It looks similar to others that we removed for bullying or harassment."*

The wording is interesting because it implies the training that has taken place with the abuse detection – the platform has a corpus of abusive statements which is used for training data. However, in this case, its identification is less than perfect, as this was a post about Plymouth Hoe, a location on the seafront in the city of Plymouth. Clearly, the algorithm thought I was making use of the term "Plymouth Hoe" to refer to a person rather than a place, hence the warning. Thankfully this no longer happens, and you are free to post as much as you like about Plymouth Hoe, but it is a nice vignette to illustrate the challenges of automated moderation.

Saying "Stop racism on your platform" might be an easy thing to say, to automate this is far more of a challenge.

In the second example, I refer to the following from a debate about the soon-to-be-enacted Online Safety Bill, in early September 2023[8]:

> *Importantly, user-to-user providers, as well as dedicated adult sites, will now be explicitly required to use highly effective age verification tools to prevent children from accessing them. The wording "highly effective" is crucial, because porn is porn wherever it is found, whether on Twitter, which, as my right hon. Friend the Member for Chelmsford said, is the most likely place for children to find pornography, or on dedicated adult sites.*

Age Verification (AV) has been a goal of the UK government for well over ten years now, particularly around the prevention of access to pornography by young people. The solution to this, we are told, is for pornography providers to implement age gates so any user proves they are over the age of 18 if they are in the UK and trying to access the service. There are functional challenges in basic age verification in that there is no mandatory requirement to carry a consistent form of identification with a date of birth in the UK. While there

[7] https://hansard.parliament.uk/Commons/2021-07-14/debates/E0C07F8B-EE53-42B1-AEDE-1AA8CBFEFD4B/Engagements#contribution-B1D273CB-B811-4750-A266-82366CB39CB9 [Accessed September 2023]

[8] https://hansard.parliament.uk/Commons/2023-09-12/debates/81853BB7-375E-45C0-8C9D-4169AC36DD12/OnlineSafetyBill# [Accessed September 2023]

are some that are frequently used, such as passports and driving licenses, there is no requirement to have either, and we might observe that both have a financial barrier to them.

It has been suggested (for example in the Age-Appropriate Design Code by the Information Commissioner[9]) that AI might be used for Age Estimation (AE) instead – a system when the end user appears in front of a webcam and an automated system determines whether they look over the age of 18. This is, on the face of it, a simple narrow domain where a lot of training data can be built – we simply need a corpus of those over the age of 18, and those who are not. However, regardless of the privacy challenges in collecting this data consensually, if it is difficult for a human to determine whether someone is over the age of 18 simply by appearance (which is why young people will be asked for ID when buying alcohol and that many supermarkets set the threshold to 25), how might we imagine an algorithm might perform?

Yoti[10], arguably the market leader in this field, produced a white paper last year[11] on their age estimation technology. In it, they claimed:

> *Our True Positive Rate (TPR) for 13-17-year-olds being correctly estimated as under 23 is 99.65%. This gives regulators a very high level of confidence that nobody underage will be able to access adult content. Our TPR for 6-11-year-olds being correctly estimated as under 13 is 98.91%. Our solution is configurable to meet any regulations that require prior consent before age estimation is used.*

Clearly, these are impressive statistics, and they continue to improve. But the threshold is set to identify under the age of 23, not 18. Therefore, if such systems were implemented as "solutions" we would expect some young adults who are entitled to access this legal content to be rejected, or to be asked to provide some other form of proof. Which they might not have because there is no legal requirement to do so.

This is for one use case – the prevention of access for children to pornography. There are even more complex challenges from AE systems set for 13-year-olds (the "digital age of consent" in the UK). AV/AE systems are improving all of the time, but not perfect and it is worrying when these systems are proposed as the "solution". While they will undoubtedly provide challenge to access pornography and will prevent the accidental discovery of pornography on browsers, it is not a perfectly solution and it will not stop all children from accessing this sort of content. There are also challenges in the many routes to explicit content, and, as young people have already told me, the use of privacy enhancing technologies such as Virtual Private Networks as a means of bypassing these systems that AI will not solve.

Finally, we consider a perennial challenge in this area – the sending of self-produced intimate images by minors. In an Online Safety Bill debate in December 2022[12], this statement was made:

> *All the main internet providers now have technology that can identify a nude image. It would be possible to require them to prevent nude images from being shared when, because of extended age-verification abilities, they know that the user is a child.*

Again, this is a simple statement of fact that does not stand up to scrutiny. We have discussed age verification already, but it should also be acknowledged that mobile providers do not always know the age of the user of a device on their network, particularly in the case of hand me down phones and those not bound to a contract. There are also challenges in preventing transmission, given there are so many ways to send an image from one device to another now. Is the proposal to prevent the image from being taken in the first place, or from it being sent? And if it is in the sending, is the expectation that any app with the capability to transmit an image needs this functionality?

[9] Page 34 of the code makes explicit reference to using AI for age estimation. https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-data-protection-themes/age-appropriate-design-a-code-of-practice-for-online-services-2-1.pdf [Accessed September 2023]

[10] https://www.yoti.com/ [Accessed September 2023]

[11] https://www.yoti.com/wp-content/uploads/Yoti-Age-Estimation-White-Paper-May-2022.pdf [Accessed September 2023]

[12] https://hansard.parliament.uk/Commons/2022-12-05/debates/E155684B-DEB0-43B4-BC76-BF53FEE8086A/OnlineSafetyBill (Column 112) [Accessed September 2023]

Nudity detection, of itself, remains a challenge (Samal et al., 2023). A corpus of tumours might have a similar format of medical image with associated metadata that states this is a tumour and its nature. For nudity, this is far more of a challenge given the diversity of images than could fall into this category. And this is compounded further in that the problem domain is to identify self-produced intimate images of minors (given it is perfectly legal for an adult to take and send an intimate image to another adult). If we are to consider the training data needed for this, we can see that there are significant ethical and legal challenges in achieving this, particularly for a private company. Again, while the statement might be simple, the reality of implementation is far more complex.

So, to return to our question: Can AI Protect Children Online?

No, not on its own.

The environments in which online harms occur are large and complex, exactly the sort of environments that challenge machine-learning approaches. That is not to say (and this is important to state because in these debates many expect a binary perspective rather than one that tries to understand the complexities of the issues and bring nuance to the debates) that technology does not have a part to play in keeping children safe online.

Clearly, there are tools that make use of machine learning that will help, and we should not let the perfect be the enemy of the good. Age verification/estimation is a perfectly acceptable tool to challenge easy access to pornography. But it should not be viewed as a global solution, and, more importantly, we should not say "Technology should solve this, so we don't need to do anything else". As unpalatable as it is, we need to acknowledge that even with technical barriers in place some young people will still access pornography, and some will be upset by what they see. So, we need to have other measures in place other than an assumption that the provider should stop it all, particularly as the "provider" in some use cases will be a peer.

Furthermore, platforms are already using machine learning systems to try to detect abuse and raise warnings. However, again, they cannot be viewed as complete solutions. There are other tools available that tend to get less coverage because they don't fit with the AI hype wave but are excellent applications of established technology supporting victims of abuse.

The StopNCII approach[13], developed by the charity SWGfL[14], with a significant in-kind contribution from Meta, does not look to automate the detection of an image but instead allows the owner of an image to prevent it from being shared. Developed to support adult victims of image-based abuse, but also used by NCMEC[15] in the US with young people, the system uses the well-established technique of image hashing[16], which processes the data in an image to produce a unique identifier for it, often compared to a fingerprint of the image, which they can then upload to the system, which then shares its database of hashed images with major user-to-user service providers. Therefore, if someone tries to share this image to abuse, it will not be posted. While, again, the technology is not perfect, it is an excellent example of using a simpler technical approach to empower victims. It is also a good example of a service provider and civil society working together for the public good.

However, a fundamental challenge in the online harms area is that there are many who believe there is a solution. Why can't we simply stop this sort of thing from happening? I am reminded of the rather famous (among technologists), Ranum's law, stated by cybersecurity researcher Marcus Ranum:

You Can't Solve Social Problems with Software

The challenge for technical interventions in this area is that most harm is caused by other people in, albeit online, social settings. Which, as with other areas of social policy where prohibition has been a utopian goal, will generally be more successful with a mix of intervention, some technical, some educational and some drawing from public health, rather than assuming some clever technology can stop it, just because it occurs

[13] https:/www.stopncii.org/ [Accessed September 2023]
[14] https://www.swgfl.org.uk/ [Accessed September 2023]
[15] https://takeitdown.ncmec.org/ [Accessed September 2023]
[16] https://stopncii.org/how-it-works/ [Accessed September 2023]

online. To finish I often use this quote, from a thirteen-year-old young person I was speaking to about what we mean by "online safety". Their view was not one of technical intervention or prohibition, it was far more holistic:

> *Is it when you know who to tell if you're upset by something that happens online?*

There are lots of ways technology can help, but none will be *the* solution.

© Professor Andy Phippen 2023

References and Further Reading

Kautz, H., 2022. The third AI summer: AAAI Robert S. Engelmore memorial lecture. AI Magazine, 43(1), pp.105-125.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A., 2021. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), pp.1-35.

O'Neil, C. Weapons of math destruction: How big data increases inequality and threatens democracy. 2016. Broadway Books.

Samal, S., Nayak, R., Jena, S. et al. Obscene image detection using transfer learning and feature fusion. 2023. Multimed Tools Appl 82, pp. 28739–28767.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D. and Chi, E.H., 2022. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

Zhang, L., Wang, S., & Liu, B. Deep learning for sentiment analysis: A survey. 2018. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.

© Professor Andy Phippen 2023