Can AI Protect Children Online?

Andy Phippen
aphippen1@bournemouth.ac.uk

Agenda

- Can AI be used to tackle issues related to online harms?
- The policy rhetoric
- Are there specific challenges related to online harms that present greater challenges to code?
- Do they introduce ethical and rights challenges?
- Are there alternatives?

What we are not going to do:
A deep dive into the workings of machine learning systems

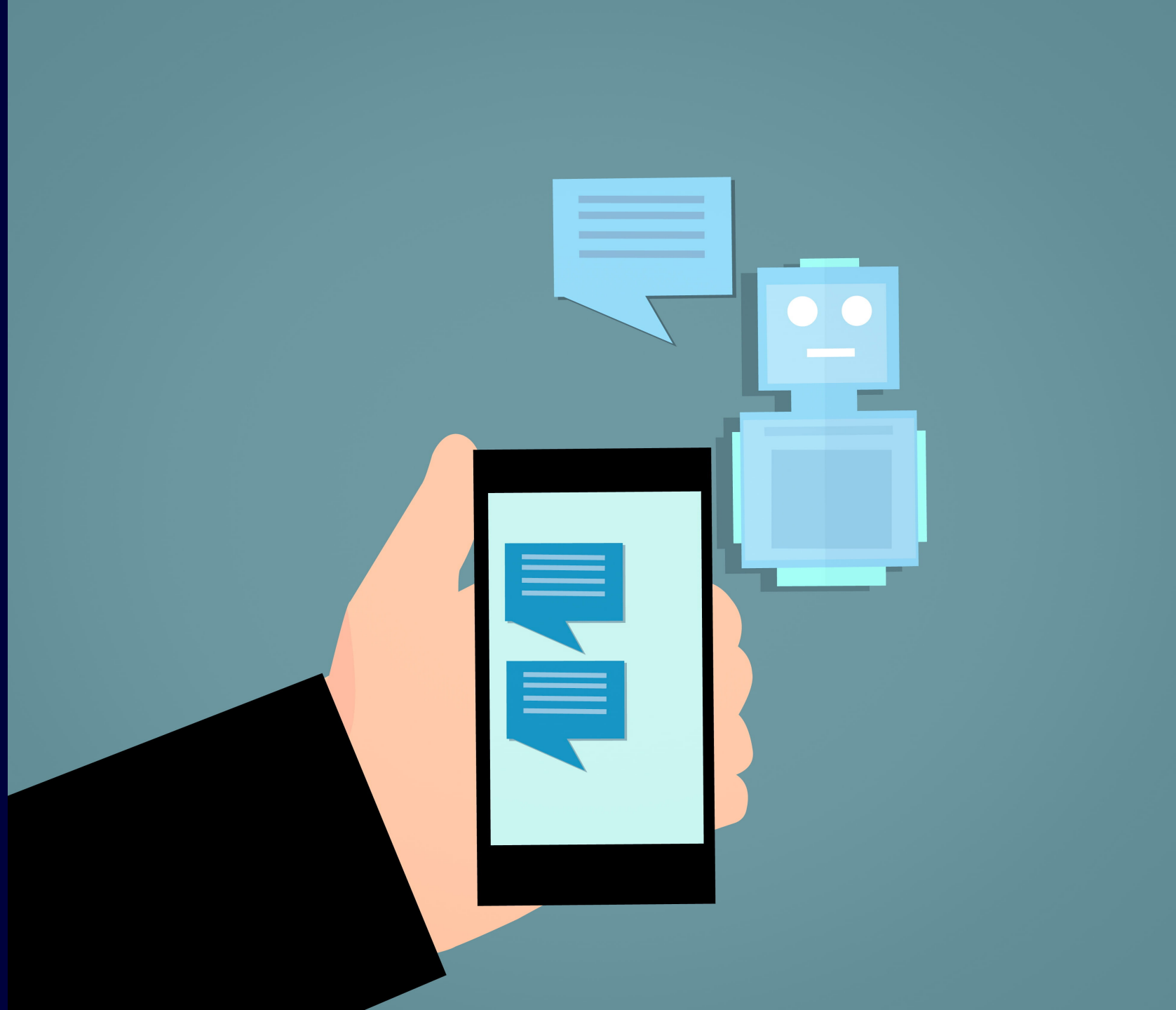
About me…

- Started career in an AI research lab in the 90s
- Moved more and more into tech ethics and social responsibility
- Specialisms around young people and online harms for the last 20 years
- Working across tech, law, society and politics on these issues
- Fellow of the BCS
- A parent

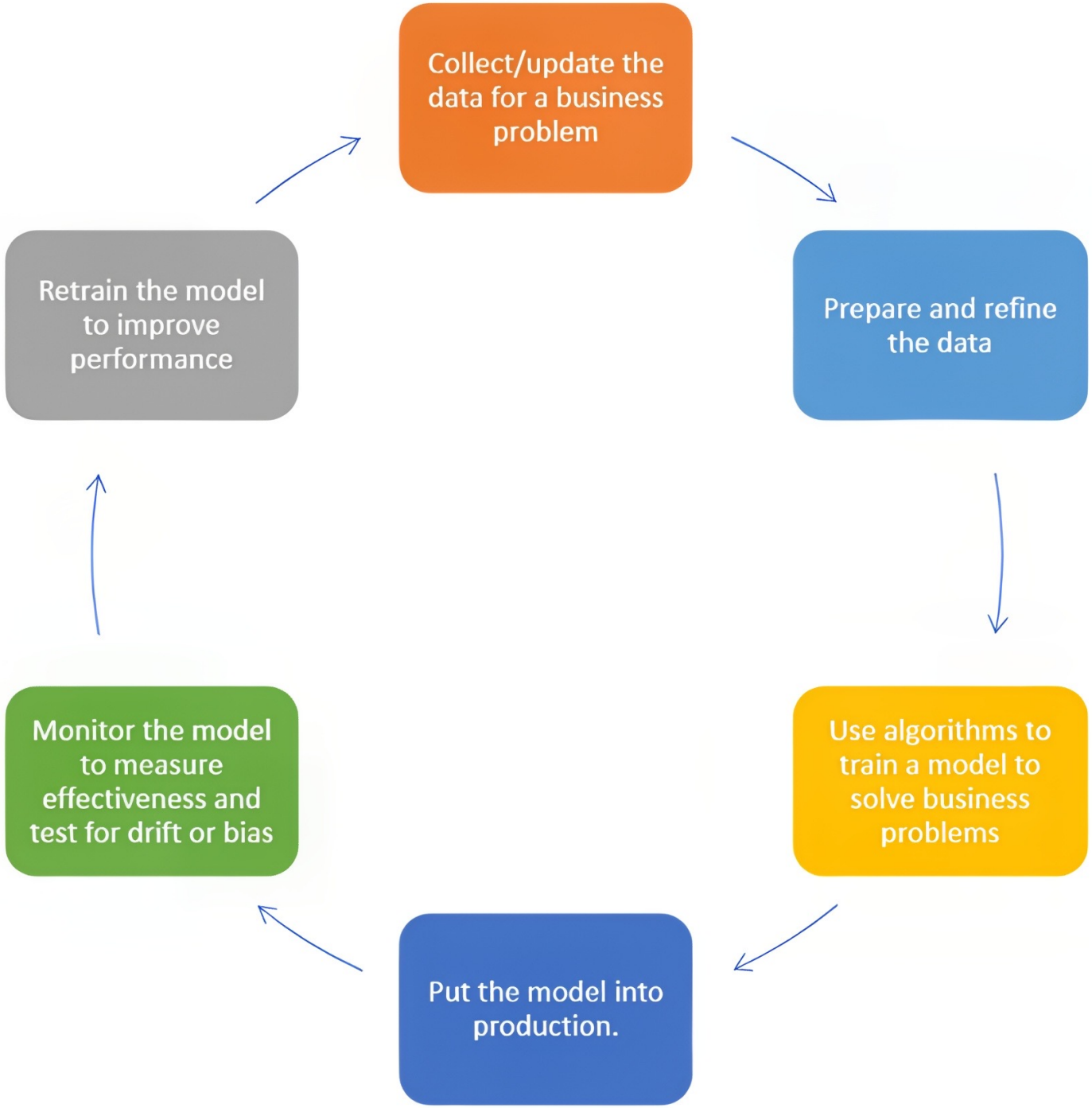
Sternenjaeger, CC0,
via Wikimedia Commons

In general

- Massive improvements in clearly defined tasks in narrow domains
- Messy and complex tasks which have less clear system boundaries are still more difficult
- Still challenges in broad understanding of the world, “common sense”, inference from small data sets, creativity…



Piqsels, CC0,
via Wikimedia Commons



“A-Levels and GCSEs: Boris Johnson blames ‘mutant algorithm’ for exam fiasco”

“UK passport photo checker shows bias against dark-skinned women”

Jamie Zawinski, Attribution, via Wikimedia Commons
https://upload.wikimedia.org/wikipedia/commons/9/9b/The.Matrix.glmatrix.2.png

Tackling Online Harms with Algorithms – The Political Discourse

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

Your Comment May Go Against Our Community Standards

It looks similar to others that we removed for bullying or harassment.

Plymouth: Plymouth Hoe by Lewis Clarke, CC BY-SA 2.0 <https://creativecommons.org/licenses/by-sa/2.0>, via Wikimedia Commons

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 accessing them. The wording "highly effective" is crucial, because porn is porn wherever it is found…

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.

British Board of Film Classification, Public domain, via Wikimedia Commons

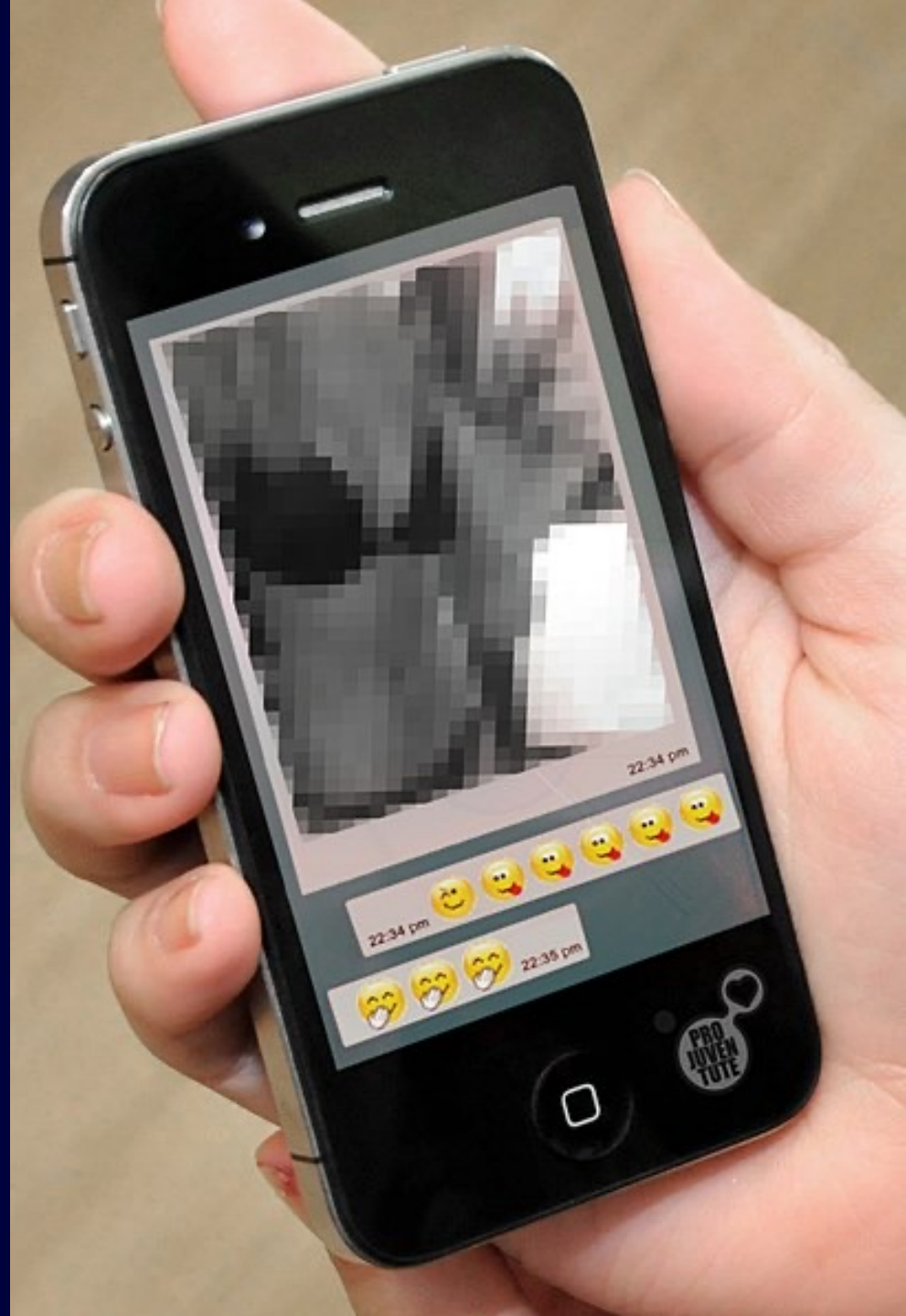
All the main internet providers now have technology that can identify a nude image. It would be possible to require them to prevent nude images from being shared when, because of extended age-verification abilities, they know that the user is a child.

Detecting a nude image

Detecting a nude image of a minor

Determining the age of the device user

Preventing transmission



Can AI Protect Children Online?

No.

But it can help as part of a toolbox of technical interventions.

StopNCII.org

Stop Non-Consensual Intimate Image Abuse

Privacy Policy | Create Your Case | English

Check Case Status | Resources and Support | How It Works | About Us | Partners | FAQ

What do you do if someone is **threatening to share** your intimate images?

Create Your Case


sad
news
dead
inappropriate
roblox
animal
friend
comments
upset
anything
dog
racist
friends
youtube
people
online
getting
stuff
picture
rude
abuse
racism
nasty
pictures
things
something
called
seen
person
died
someone
messages
scary
made
bad
really
thing
mean
video
videos
momo
saw
one
upsetting
swearing
animals
just
bullying
game
saying

You Can’t Solve Social Problems with Software

Ranum’s Law

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