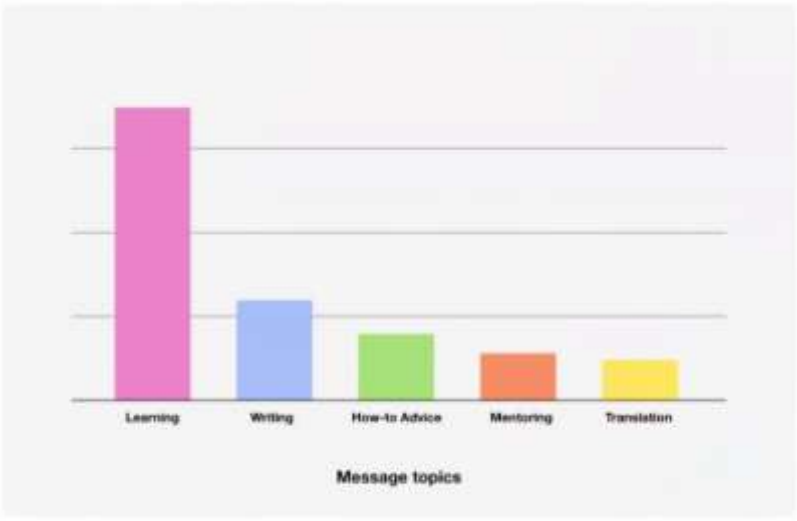
Meeting with OpenAI

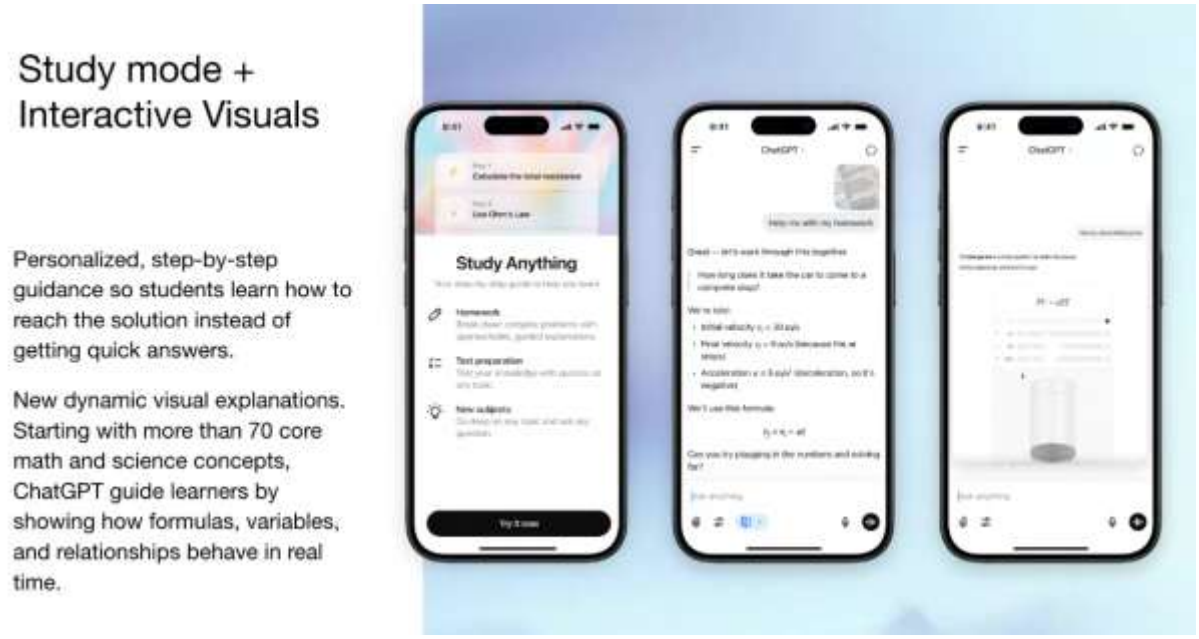
1. Julie Darger	Denmark
2. Hugo Krief	France
3. Annekathrin Cornelius	Germany
4. Frieda Groschup	Germany
5. Grete Raidma	INHOPE
6. Fiona Jennings	Ireland
7. Jane McGarrigle	Ireland
8. Catriona Mulcahy	Ireland
9. Liudas Mikalkevičius	Lithuania
10. Rūta Žiogelė	Lithuania
11. Zoé Scherer	Luxembourg
12. Davinia Marie Muscat	Malta
13. Vineeca Kuo	The Netherlands
14. Simona Levi	North Macedonia
15. Patrícia São João	Portugal
16. Marta Cantón Martínez	Spain
17. Rafa Nicolazzi	OpenAI
18. Caitlin Niedermeyer	OpenAI
19. Karl Hopwood	Insafe/EUN
20. Nayia Stavrou	
21. Pilar	

Learning is the most common reason for young people to use ChatGPT.



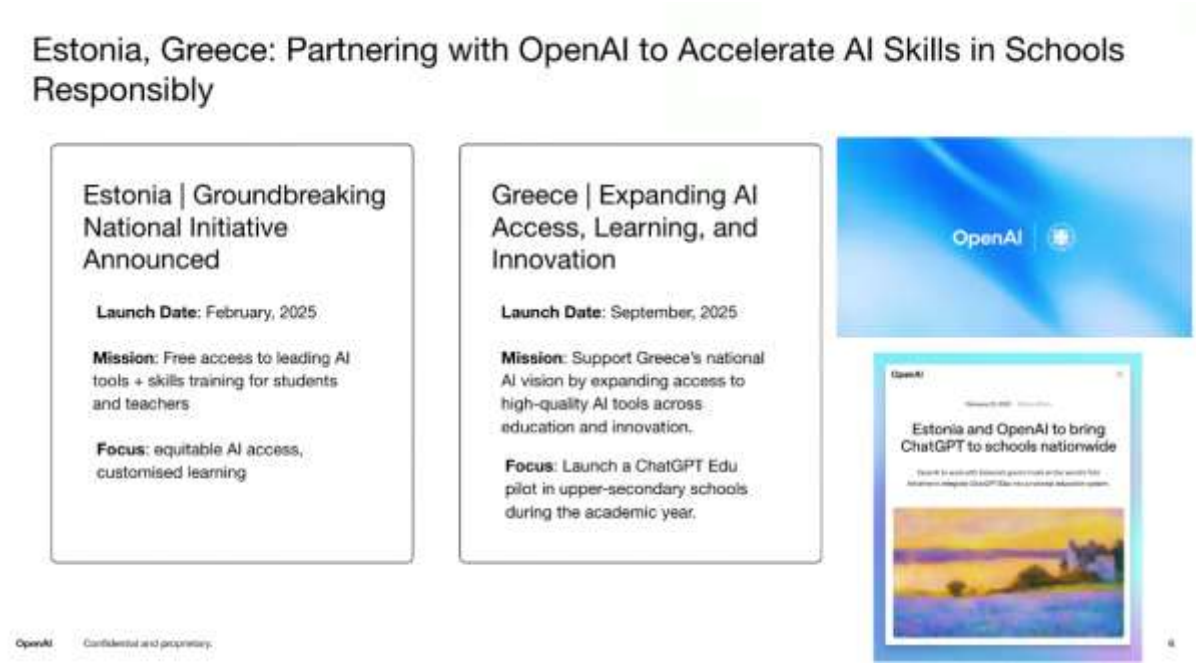
There are around 900 million weekly active users making ChatGPT the number one learning platform in the world. There is significant investment in features to reflect the demand.

Study mode will guide a user to think about the answer, they will not be given a direct answer automatically with the goal to be better for the development of cognitive ability.



There is also an aim to make learning more interactive, particularly with maths or science where the use of images is useful. Initially around 70 core maths and science concepts have been created with images.

There are several countries that are partnering with OpenAI to accelerate the development of responsible AI skills in schools e.g. Estonia and Greece. The partnerships are with the government as they are able to reach schools more easily and know how to bring AI to them.



ChatGPT Edu is being used as an AI tutor and initial findings show that it has a particular impact with STEM subjects, academic writing and professional skills. The preliminary findings are positive and show that users have better outcomes, better motivation and improved high order thinking.

Meta-Analysis Overview
Source: Wang & Fan (2025), Humanities & Social Sciences Communications
Scope: 51 experimental studies (Nov 2022 – Feb 2025)

Key Findings

- **+86%** improvement in learning performance measured by test scores and academic outcomes
- **+45%** Boost in learning perception (g = 0.456) measured by motivation, clarity, and student confidence
- **+46%** Improvement in higher-order thinking (g = 0.457) measured by growth in critical thinking, creativity, metacognition

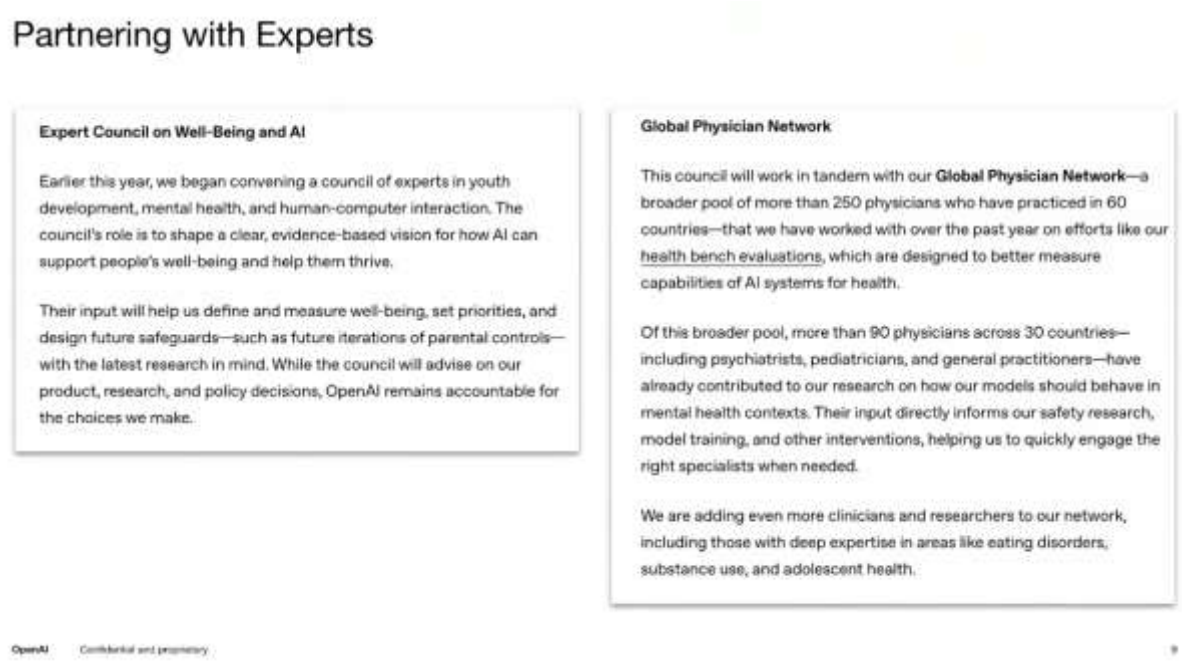
When ChatGPT Works Best

- STEM, academic writing, professional skills
- Highest impact when used as an intelligent tutor

ChatGPT Edu is more than a chatbot—it's a flexible, proven, AI-powered tutor that scales across disciplines and boosts performance, engagement, and thinking skills.

A project is underway with the University of Tartu in Estonia which is developing and validating a learning management measurement suite looking at how AI can affect learning over time. The study is involving over 20,000 students from 16-18 – this goes beyond test scores and looks at cognitive ability.

Open AI has an expert council on well-being and AI and a global protection network (see below).



The council guides OpenAI's ongoing work and involves leading researchers and experts, considering the impact that AI can have on emotion, motivation and mental health as well as what healthy interactions should look like. There is also a broader pool of mental health clinicians involved in the global physician network. The group is currently looking at how to notify parents if their teen is in distress.

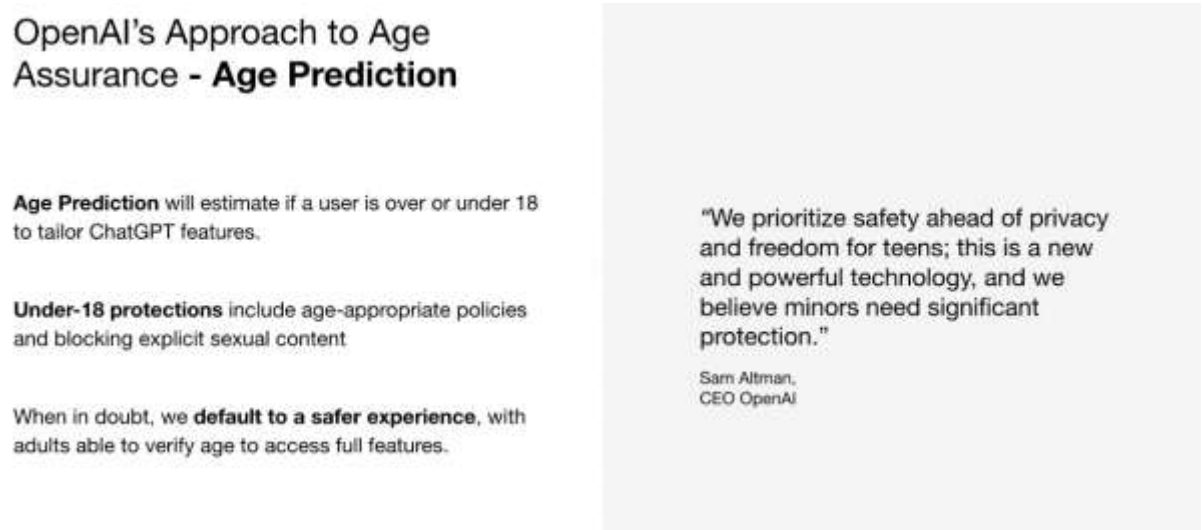
It is important for OpenAI to know who is using their services in order to be able to ensure that teens have age-appropriate protections in place by default as well as providing tools that parents can deploy. Age assurance is a term which spans a range of different ways for being able to understand age.

Age declaration means that a user self-declares their age – usually by providing a date of birth.

Age prediction is where the platform will use a range of behavioural signals in order to determine age. This will include language, activity, content that is being requested etc. and device settings.

Age verification means that a user will prove their age, often by uploading some government approved ID.

If it is not possible to have a good estimation of age then the account will be automatically defaulted into teen settings. This is not currently available in the EU yet but will be soon.



Age prediction is designed to reduce exposure to sensitive content such as graphic violence, graphic challenges and depictions of self-harm. This approach is rooted in the science of child development around impulse controls, peer influence, emotional regulation. Accounts will always default to a safer experience if there is any uncertainty about age. There is a right to appeal and users will have an option to confirm their age through age verification.

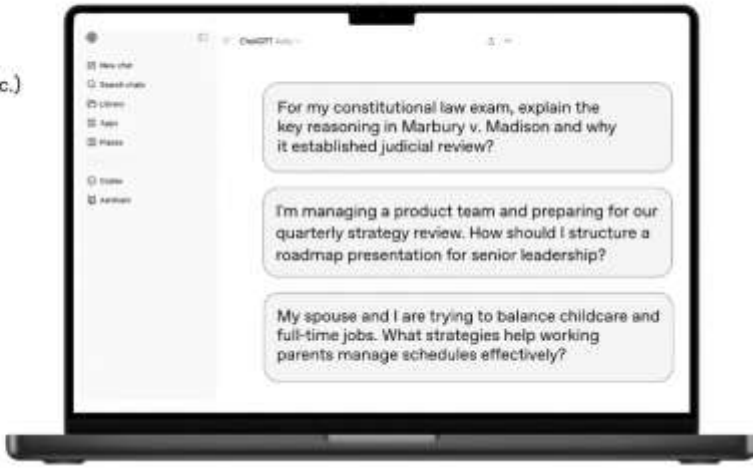
Signals to predict U18

Content signals

- Writing style (emoji, slang, etc.)
- Topics

Behavioral signals

- Time of day
- Message and conversation counts



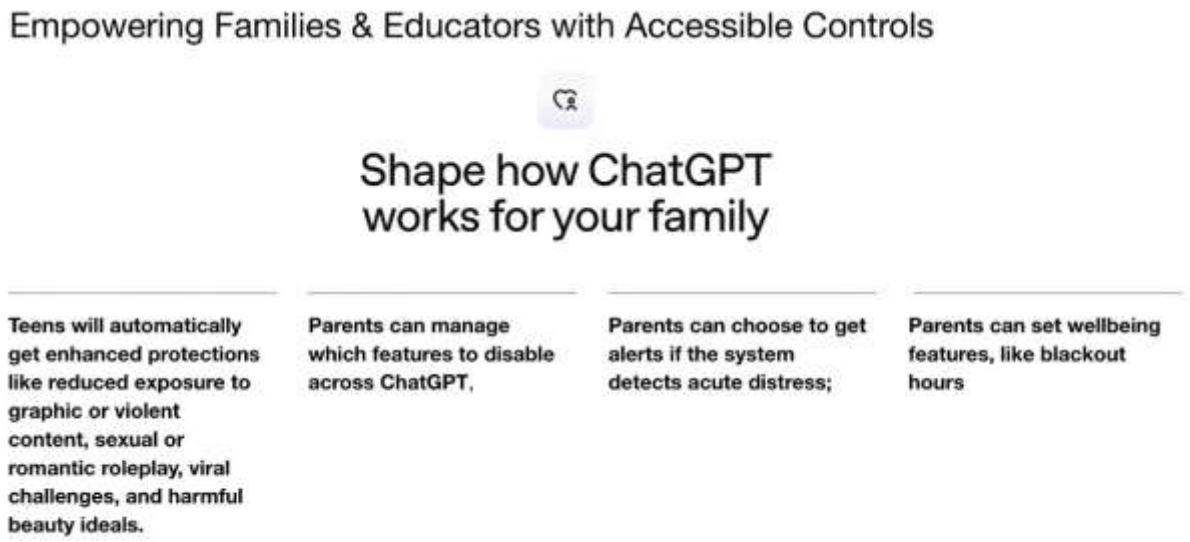
A model spec exists and has been adapted with specific principles adapted for under 18s. Consideration is given as to how the model should behave differently for under 18s. These principles are anchored in four guiding commitments.

- Put teen safety first, even when it may conflict with other goals
- Promote real-world support by encouraging offline relationships and trusted resources
- Treat teens like teens, neither condescending to them nor treating them as adults
- Be transparent by setting clear expectations

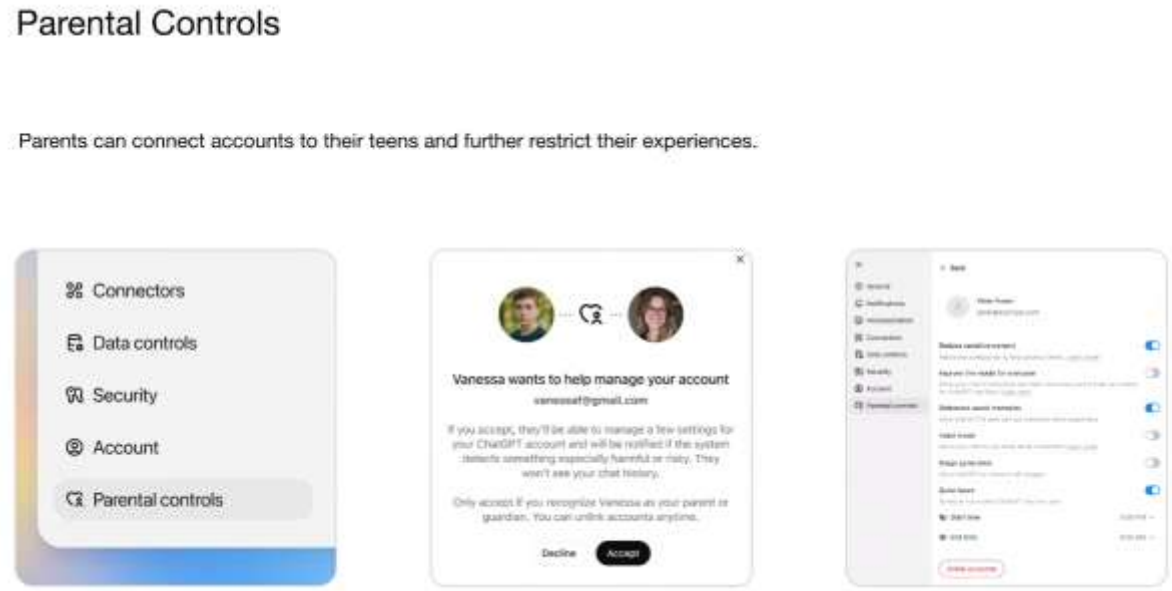
There is a road map for how OpenAI want the model to interact with users and how it should respond to specific issues. This is a layer on top of the model spec.

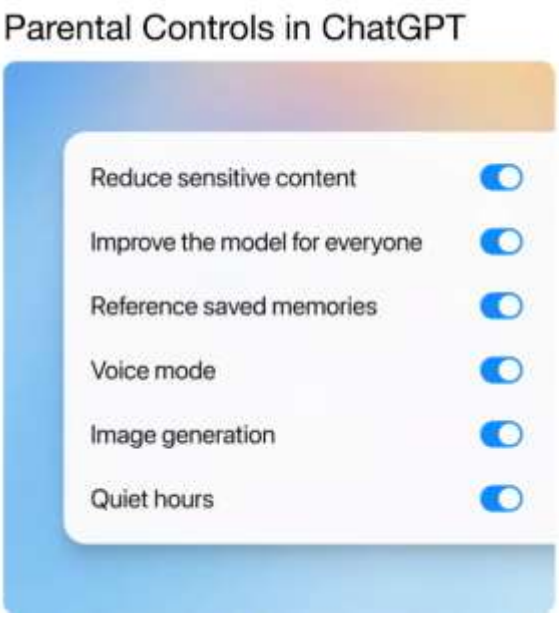
Teen policies

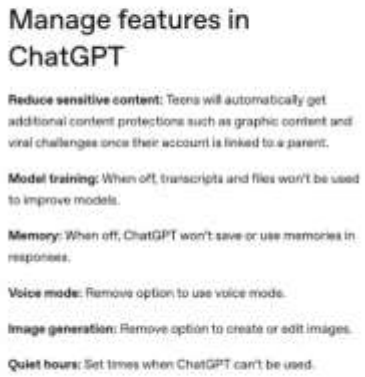
Policy	U18 specific
Self-harm	For all users, it's prohibited to romanticize or provide instructions on self-harm or suicide For U18, this **boundary exist even if the context is fictional, hypothetical, historical, or educational.**
Romantic or erotic roleplay	For all users, it's prohibited to engage in role-play that could undermine real-world ties. For U18 users, the assistant **additionally cannot engage in immersive romantic roleplay, first-person intimacy, or pairing the assistant romantically with a teen—even if a similar scene would be allowed between consenting adults.**
Graphic or explicit detail	For all users, there are limits to gore and explicit sexual or violent details. For U18, This **boundary exists even for educational discussions**. Also, the assistant should not enable first-person sexual or violent roleplay even if it is non-graphic and non-explicit.
Dangerous activities and substances	For all users, there are restrictions to actionable instructions for harmful and unlawful acts For U18, it **also covers activities that may be legal for adults but pose heightened risk to adolescents, including age-restricted challenges, stunts, or risky behaviors.**
Body image and disordered eating	For all users, there is no encouragement to unhealthy eating behaviors. For U18, there is an **extra care to not enable any appearance critiques, image comparisons, gendered appearance ideals or restrictive eating advice (even when such content may be acceptable for adults, for example, intermittent fasting).**
Defaulting to safety	For adults, there is a balance between autonomy and safety For U18, **the assistant errs further on the side of safety over autonomy.**



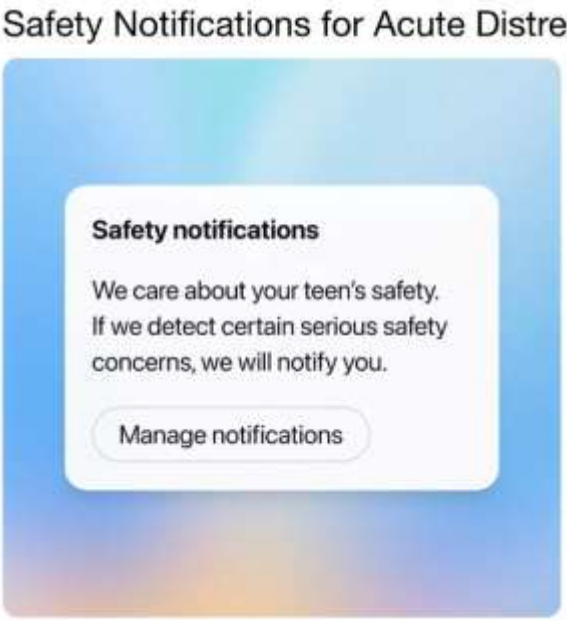
Parental controls are offered to parents in order to shape how chat works for teens in their care. A parent/carer can connect their account to their teen's account and can receive alerts if the platform detects distress. Once accounts are linked the default is that a parent will be signed up to receive alerts and will need to opt out of this to stop it. The alerts can be received in a variety of ways, call, email, text.

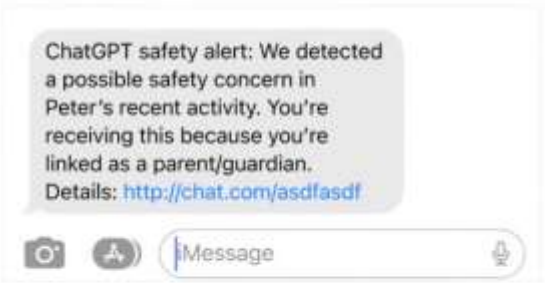






Once a parent links their account to their teens they can manage settings. If a parent sets a specific setting the teen can't change the setting once the accounts are linked.





So far there has been good engagement from families but OpenAI is keen to amplify more with parents. A trusted contact feature is being rolled out soon. The alert will be available for adults too.

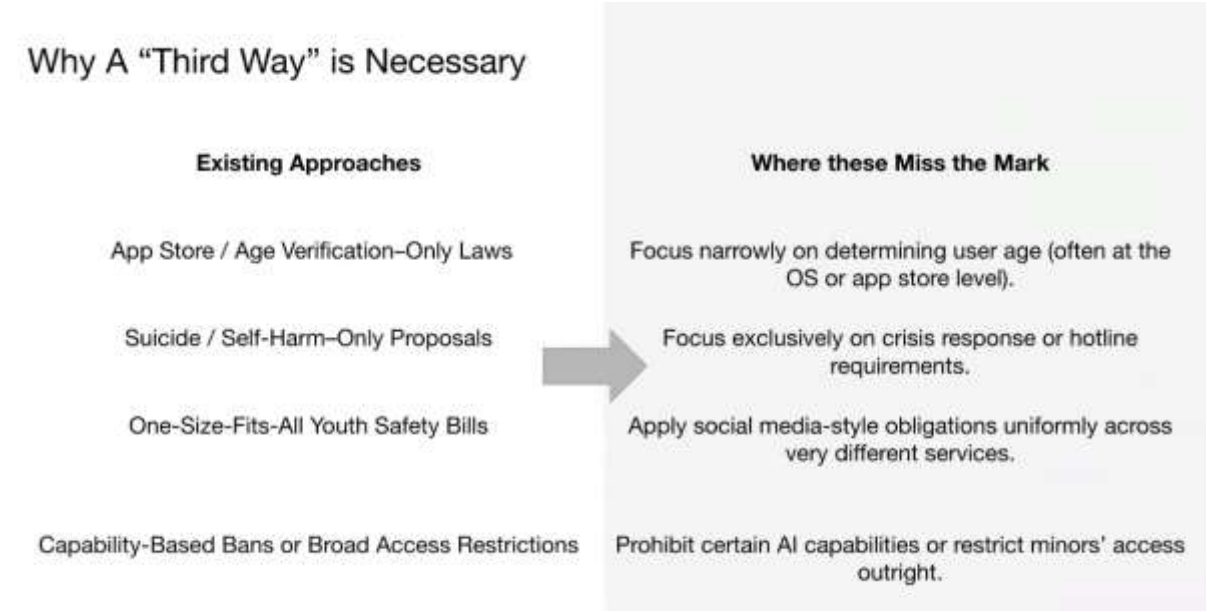
A family guide has been created to help teens use AI responsibly. This includes tips for parents to talk to their children about AI and how to empower them – it considers what ChatGPT can and cannot do.

OpenAI is working with Common Sense Media to develop Parent and Kids Safe AI Act – this will be the first framework for protecting children in the age of AI and uses a similar risk assessment framework to that which the DSA leverages.

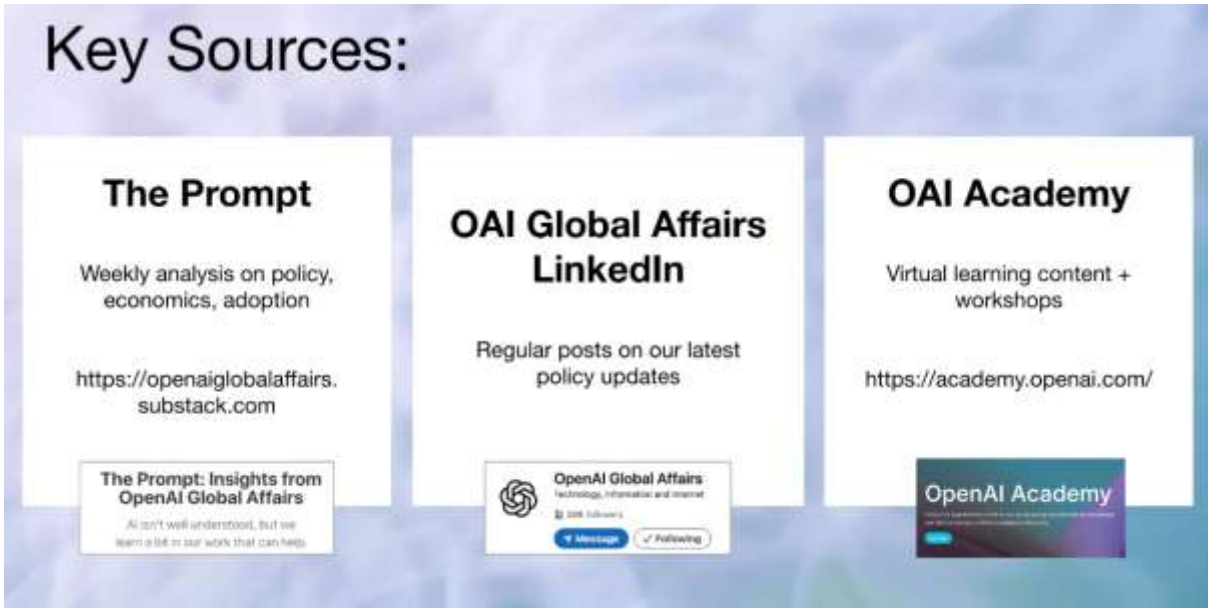
The idea is that this is a risk based and privacy preserving framework, recognising that AI has unique opportunities and risks. The framework will set clear and enforceable recommendations.

1. Need to know when a user is a minor – so need to implement age verification and then default to safer if unsure.
2. Assess risk each year – so future proof as risk changes – requirement for provider to document how risks are mitigated – including self-harm – need to have a child safety policy
3. Concrete product safeguard – crisis response protocol – manipulative and deceptive design features. Under 18 model policy
4. Parent controls
5. Independent assessment – hold products accountable – have to have an annual child safety audit – submit to attorney general.

The framework has been announced in the US and OpenAI is looking to see how this can be adapted in Europe and other markets.



Some proposals were too narrow – hence doing the above – wanted to broaden this out.



An OpenAI academy will showcase best practice on how to use ChatGPT and AI more broadly.

Q&A

Is model training turned off by default?

It depends – we note differences – sensitive content is on by default but a parent could turn it off. The aim is to improve the model for everyone so model training is on by default – but it can be turned off. This is the same with the other settings (apart from sensitive content), they are all on by default.

If acute harm is identified – what support is the teen given?

For a teen – with acute harm identified – or a signal that they're trying to self-harm – we will signpost them to a helpline and resources etc. OpenAI uses Throughline to identify the most appropriate helplines.

In learning mode what happens if a teen takes a picture of a task and say can you do it for me? How does it work – how can you prevent that ChatGPT does everything? Can this be forced on?

If ChatGPT is being used in school then teachers need to guide how it is leveraged and ideally use stud mode. https://openai.com/index/chatgpt-study-mode/ Teachers are best placed to do this.

https://openai.com/index/ai-literacy-resources-for-teens-and-parents/