
winder.AI
Online Safety and MLOps
Dr. Phil Winder, CEO
goto;
GOTO
AMSTERDAM 2023
#GOTOams

goto;

GOTO
Guide

LET US HELP YOU

Ask questions through **the app**

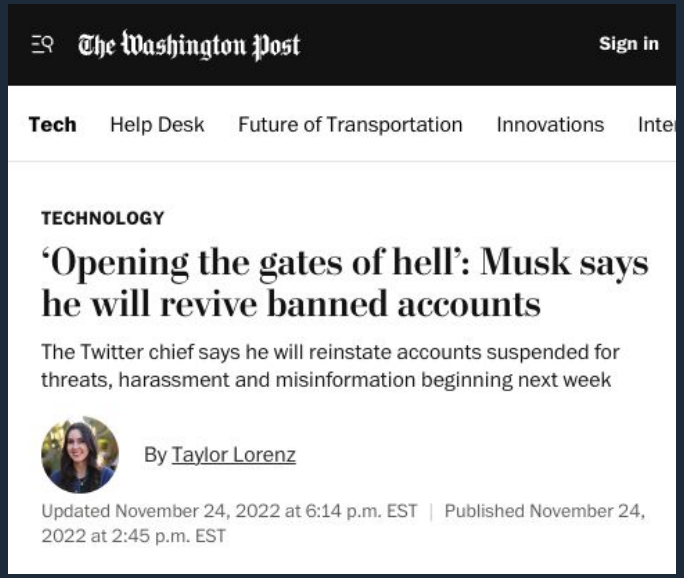
★★★★★ *with 11 stars*

also remember to rate session

THANK YOU!

#GOTOams

Social Media



The Washington Post

Sign in

Tech Help Desk Future of Transportation Innovations Inte

TECHNOLOGY

‘Opening the gates of hell’: Musk says he will revive banned accounts

The Twitter chief says he will reinstate accounts suspended for threats, harassment and misinformation beginning next week

By Taylor Lorenz

Updated November 24, 2022 at 6:14 p.m. EST | Published November 24, 2022 at 2:45 p.m. EST

https://www.washingtonpost.com/technology/2022/11/24/twitter-musk-reverses-suspensions/

Support us

The Guardian

Newspaper of the year

News Opinion Sport Culture Lifestyle

The Guardian view Columnists Cartoons Opinion videos Letters

Opinion

This article is more than 3 months old

Elon Musk’s Twitter is fast proving that free speech at all costs is a dangerous fantasy

Nesrine Malik

Reinstating the likes of Donald Trump and Kanye West looks likely to turn the social media site into an extremist ghetto

Mon 28 Nov 2022 06.00 GMT

https://www.theguardian.com/commentisfree/2022/nov/28/elon-musk-twitter-free-speech-donald-trump-kanye-west

winder.AI

Is Speech Really Free?

ELIZABETH II

Human Rights Act 1998

1998 CHAPTER 42

ct to give further effect to rights and freedoms guaran
der the European Convention on Human Rights; to n
ovision with respect to holders of certain judicial offices
come judges of the European Court of Human Rights; an
nnected purposes. [9th November 1

https://www.legislation.gov.uk/ukpga/1998/42/schedule/1/part/I/chapter/9

Human Rights Act 1998

– Coopted from EU European Convention

Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers.

Subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary.

winder.AI

Other Laws

Public Order Act 1986

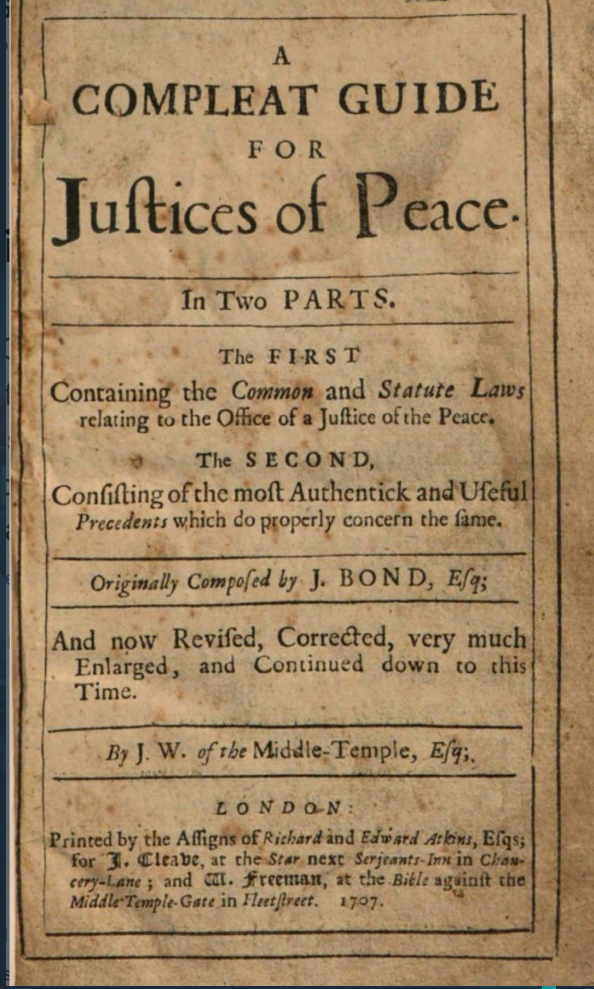
A person is guilty of an offence if he/she:

> *(a) uses threatening [or abusive] words or behaviour, or disorderly behaviour, or*
>
> *(b) displays any writing, sign or other visible representation which is threatening [or abusive],*

within the hearing or sight of a person likely to be caused harassment, alarm or distress thereby.

Justices of the Peace Act 1361

Riotous and barratous behaviour that disturbs the peace of the King


A
COMPLEAT GUIDE
FOR
Juſtices of Peace.

In Two PARTS.

The FIRST
Containing the *Common* and *Statute Laws* relating to the Office of a Juſtice of the Peace.

The SECOND,
Conſiſting of the moſt Authentick and Uſeful *Precedents* which do properly concern the ſame.

Originally Compoſed by J. BOND, *Eſq;*

And now Reviſed, Corrected, very much Enlarged, and Continued down to this Time.

By J. W. *of the* Middle-Temple, *Eſq;*

LONDON:
Printed by the Aſſigns of *Richard* and *Edward Atkins*, Eſqs; for J. Cleave, at the *Star* next *Serjeants-Inn* in *Chancery-Lane*; and M. Freeman, at the *Bible* againſt the *Middle-Temple-Gate* in *Fleetſtreet*. 1707.

winder.AI

Social Media

Elon Musk ✓
@elonmusk

Replying to @unusual_whales

The note is incorrect. There are ~2300 active, working employees at Twitter.

There are still hundreds of employees working on trust & safety, along with several thousand contractors.

Less than 10 people from my other companies are working at Twitter.

7:57 AM · Jan 21, 2023 · **2.3M** Views

1,982 Retweets **171** Quotes **24.4K** Likes

Tip

https://twitter.com/elonmusk/status/1616706530841333761

https://twitter.com/yoyoel/status/1588657228462317568

Yoel Roth ✓
@yoyoel

Yesterday's reduction in force affected approximately 15% of our Trust & Safety organization (as opposed to approximately 50% cuts company-wide), with our front-line moderation staff experiencing the least impact.

10:19 PM · Nov 4, 2022

467 Retweets **222** Quotes **4,457** Likes

Tip

Common

BBC

NEWS

Menu

Technology

Twitter sued over antisemitic posts left online

25 January

Josephine Ballon and Avital Grinberg are from the groups taking action

By Zoe Kleinman
Technology editor

Twitter is being sued in Germany by two groups claiming the social network failed to remove six posts attacking Jewish people and denying the Holocaust, after they were reported.

https://www.bbc.co.uk/news/technology-64404590

Online Safety

Instagram and Pinterest used algorithms that resulted in there being "binge periods" of material, some of which was selected and provided for Molly without her having requested it.

"These binge periods are likely to have had a negative effect on Molly. Some of this content romanticised acts of self-harm by young people on themselves. Other content sought to isolate and discourage discussion with those who may have been able to help.

"It is likely that the above material viewed by Molly, already suffering with a depressive illness and vulnerable due to her age, affected her mental health in a negative way and contributed to her death in a more than minimal way."

BBC

NEWS Menu

England | Local News | Regions | London

Molly Russell: Coroner's report urges social media changes

14 October 2022

Molly Russell was aged 14 when she died in 2017

A coroner has written to social media firms and the government calling for action following the inquest into the death of schoolgirl Molly Russell.

The 14-year-old, from Harrow, ended her life in November 2017 after viewing suicide and self-harm content online.

winder.AI

https://www.bbc.com/news/uk-england-london-63073489

Online Safety

Paedophiles are using artificial intelligence (AI) technology to create and sell life-like child sexual abuse material, the BBC has found. Some are accessing the images by paying subscriptions to accounts on mainstream content-sharing sites such as Patreon.

A "pseudo image" generated by a computer which depicts child sexual abuse is treated the same as a real image and is illegal to possess, publish or transfer in the UK.

"We already ban AI-generated synthetic child exploitation material," it said, describing itself as "very proactive", with dedicated teams, technology and partnerships to "keep teens safe".

Stability AI told the BBC it "prohibits any misuse for illegal or immoral purposes across our platforms, and our policies are clear that this includes CSAM (child sexual abuse material).

"We strongly support law enforcement efforts against those who misuse our products for illegal or nefarious purposes".

winder.AI

https://www.bbc.com/news/uk-65932372

Ofsted

“feedback to the senior leadership team, inspectors had said that a boy doing a dance move akin to flossing was evidence of the sexualisation of children at the school. There were also said to be claims of child-on-child abuse, which turned out to be a playground fight.”

"This one-word judgement was just destroying 32 years of her vocation. Education was her vocation, 32 years summed up in one word 'inadequate'.”

winder.AI

https://www.bbc.co.uk/news/uk-england-berkshire-65021154

BBC

NEWS Menu

England | Local News | Regions | Berkshire

Ofsted: Head teacher's family blames death on school inspection pressure

5 minutes ago

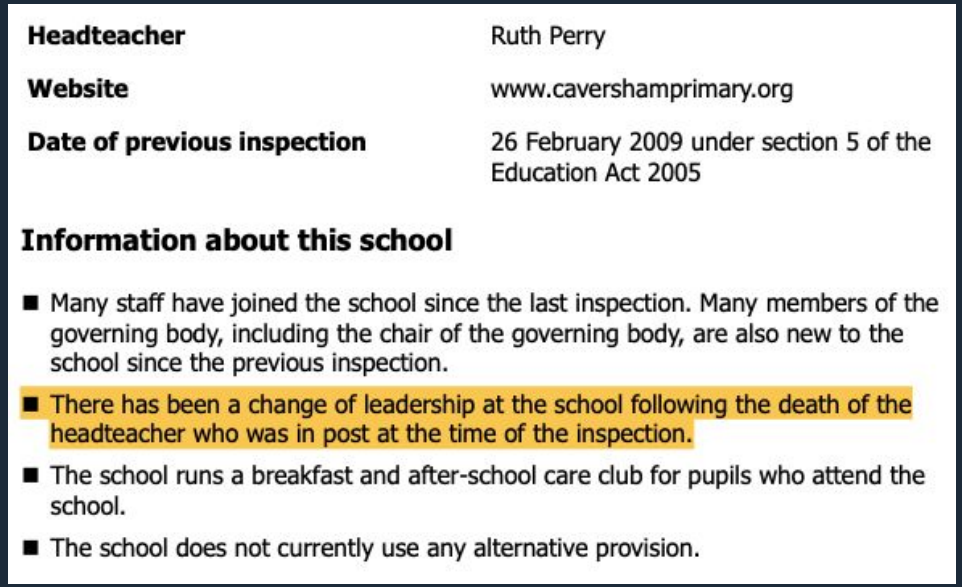
Ruth Perry was the head at Caversham Primary School in Reading

By Branwen Jeffreys & Indy Almroth-Wright & Stephen Stafford
Education editor

A head teacher who took her own life ahead of a school inspection report was under "intolerable pressure", her family has said.

Ofsted

- Irene Hogg, 58, Head, Glendinning Terrace Primary, Galashiels, suicide Mar 2008.
- Carol Woodward, 58, Head, Woodford Primary School, Plymouth, suicide Jul 2015.
- Ruth Perry, 53, Head, Caversham Primary School, Reading, suicide Jan 2023.



Headteacher	Ruth Perry
Website	www.cavershamprimary.org
Date of previous inspection	26 February 2009 under section 5 of the Education Act 2005

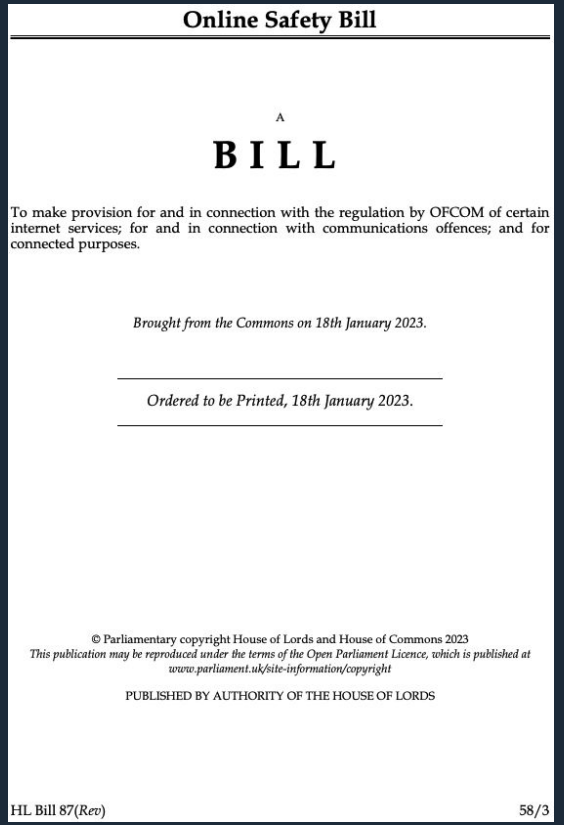
Information about this school

- Many staff have joined the school since the last inspection. Many members of the governing body, including the chair of the governing body, are also new to the school since the previous inspection.
- There has been a change of leadership at the school following the death of the headteacher who was in post at the time of the inspection.
- The school runs a breakfast and after-school care club for pupils who attend the school.
- The school does not currently use any alternative provision.

https://reports.ofsted.gov.uk/provider/21/109778

Online Safety Bill

- Remove all illegal content.
- Remove content that is banned by their own terms and conditions.
- Empower adult internet users to tailor the type of content they see.
- Children will be automatically prevented from seeing harmful content without having to change any settings.



Online Safety Bill

A

BILL

To make provision for and in connection with the regulation by OFCOM of certain internet services; for and in connection with communications offences; and for connected purposes.

Brought from the Commons on 18th January 2023.

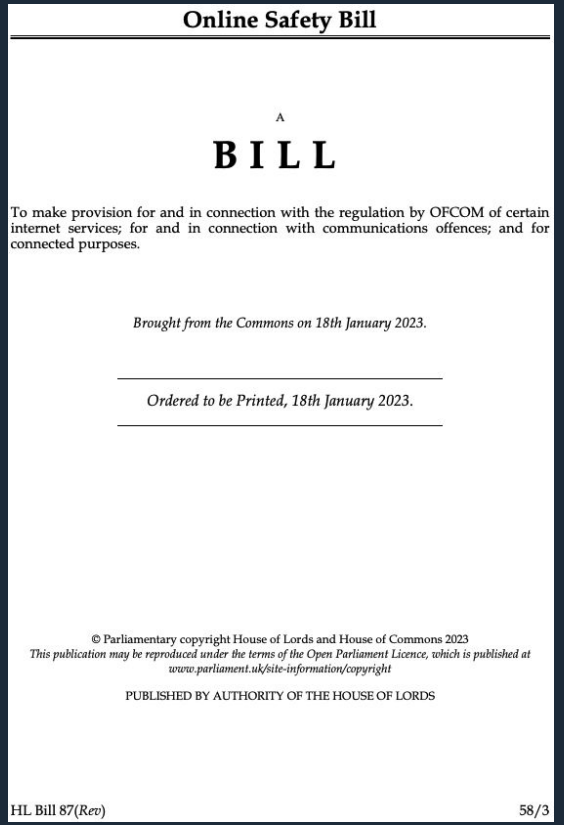
Ordered to be Printed, 18th January 2023.

PUBLISHED BY AUTHORITY OF THE HOUSE OF LORDS

HL Bill 87(*Rev*) 58/3

What is “Illegal Content”?

- child sexual abuse
- controlling or coercive behaviour
- extreme sexual violence
- fraud
- hate crime
- inciting violence
- illegal immigration and people smuggling
- promoting or facilitating suicide
- promoting self harm
- revenge porn
- selling illegal drugs or weapons
- sexual exploitation
- terrorism



Online Safety Bill

A

BILL

To make provision for and in connection with the regulation by OFCOM of certain internet services; for and in connection with communications offences; and for connected purposes.

Brought from the Commons on 18th January 2023.

Ordered to be Printed, 18th January 2023.

PUBLISHED BY AUTHORITY OF THE HOUSE OF LORDS

HL Bill 87(*Rev*) 58/3

OSB - Ramifications

- Fines up to £18 million, or **10 percent of their annual global turnover**, whichever is greater.
- Senior managers **criminally liable** if they fail to follow information requests from Ofcom.
- Ofcom will be able to require **payment providers, advertisers and internet service providers** to stop working with a site, preventing it from generating money or being accessed from the UK.
- Doesn't matter where companies are based.

Social Media - Scale

	Jul - Dec 2022	Jan - Jun 2022	Jul - Dec 2021	Jan - Jun 2021	Jul - Dec 2020
Facebook	8.7M	18.1M	14.6M	22.0M	4.0M
Twitter			0.5M	0.9M	0.2M

Jul-Dec 2022

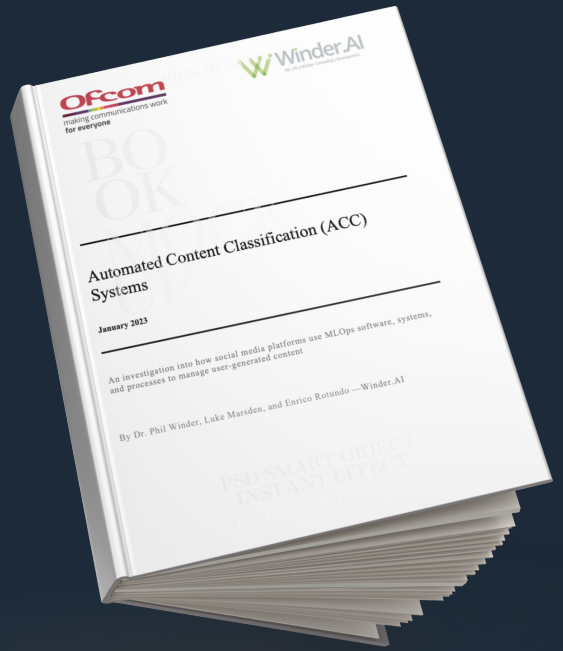
6.4B actions over all categories. 403 removals every second.

https://transparency.fb.com/data/community-standards-enforcement/suicide-and-self-injury/facebook/#content-actioned

https://transparency.twitter.com/en/reports/rules-enforcement.html

How? – AI Of Course!

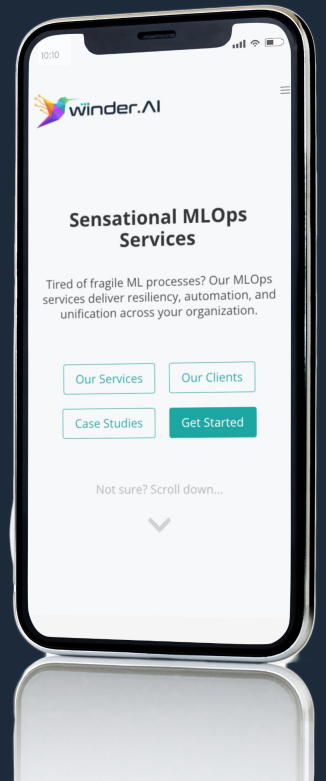
https://winder.ai/how-social-media-platforms-use-mlops-and-ai-governance-to-help-to-moderate-content/



What is MLOps?

A collection of processes and technical best practices that aim to make it easier for businesses to use AI.

- **Data management** – annotation, augmentation, registry, stores
- **Pipelining** – data, training, deployment, continuous X
- **Tracking** – registries, evaluation
- **Development** – data science lifecycle
- **Deployment** – monitoring, logging, explainability
- **Infrastructure** – capability to do work
- **Governance**
  - Provenance and lineage
  - Controls and compliance
  - Reporting
  - Auditing
  - Transparency



winder.AI

Report Methodology

- Goals
  - How do platforms build auto-moderation tools?
  - What techniques, tools, technologies, and hardware are platforms using?
  - What processes to businesses use to manage the development and operation of these tools?

Report Methodology

- Goals
  - How do platforms build auto-moderation tools?
  - What techniques, tools, technologies, and hardware are platforms using?
  - What processes to businesses use to manage the development and operation of these tools?

- Interviewed a wide range of companies
  - Large social media platforms
  - Small social platforms
  - Platforms in specific genres (e.g. adult)
  - Third-party moderation product vendors
  - Industry experts

General Findings - Data

- Both large and small businesses are working with vast amounts of data
- Data architecture is an important foundation
- Larger businesses have built bespoke data platforms, which are often open-sourced
- Improving data quality via testing and monitoring is increasingly important
- Increasingly complex annotation schemes
- Limited observability of metadata, lineage, provenance

General Findings – AI Development

- Wide range of simple and complex modelling techniques (boosting methods through to large language models)
- Common for an individual business to re-use models they are experienced with
- Development and training systems tended to align with their choice of modelling technique
- Significant development time spent altering data, not models
- Evaluation of development models is moving into production

General Findings – Deployment

- Large platforms use advanced deployment techniques. Significant amounts of experimentation in production. Dev/Stage/Prod a thing of the past for reliable AI techniques.
- Logging and monitoring is used throughout for auditing and reliability
- Large platforms have invested heavily in dedicated data centres

General Findings – AI In the Organization

- Larger platforms have more structure and distinct responsibilities
- Smaller platforms tend to be more product focused
- Most platforms leveraged third-party tools/data, also third-parties consume third-party tools/data
- Reports of initial AI integration taking > 1 year

General Findings – Costs

- AI teams are expensive, one platform suggested $10M / year.
- Moderation vendors offer a range of pricing structures:
  - Volume pricing: $1 / 1000 images, $5 / user
  - Annual pricing: $0.1 - 1M
  - Enterprise pricing: $1-10M

General Findings – Governance & Compliance

- Little distinction between corporate governance and the technical use of AI governance
- Bespoke processes that align with business' policies
- General reluctance to discuss due to perceived legal ramifications

General Findings – Challenges

- Problem definition – what is harm?
- Data retention policies – causing training to be irreproducible
- Once data comes to rest, hard to move – increasing reliance on streaming/mini-batching
- Unable to observe illegal data – AI development in the dark
- Evaluating the claims of vendors
- Transparency

General Findings – Challenges

- Problem definition – what is harm?
- Data retention policies – causing training to be irreproducible
- Once data comes to rest, hard to move – increasing reliance on streaming/mini-batching
- Unable to observe illegal data – AI development in the dark
- Evaluating the claims of vendors
- Transparency
- Hyper localization

hoisting the national flag upside down is a serious matter and under the Prevention of Insults to National Honour (Amendment) Act, 2005, this is a punishable crime and there is a provision of 3 years of imprisonment and fine.

https://www.nationalheraldindia.com/national/national-flag-hoisted-upside-down-in-jdus-patna-office-matter-reaches-high-court

EU AI Act

- Using prohibited AI technology
  - up to 40,000,000 EUR or 7% of total worldwide annual turnover
- Companies have 2 years from the enforcement date to comply with the requirements

EU AI Act – Prohibited

- Purposely manipulative
- exploits vulnerabilities
- classifying or scoring natural persons
- law enforcement (specific)
- expanding facial recognition datasets
- detecting emotions, physical or physiological features such as facial expressions, movements, pulse frequency or voice

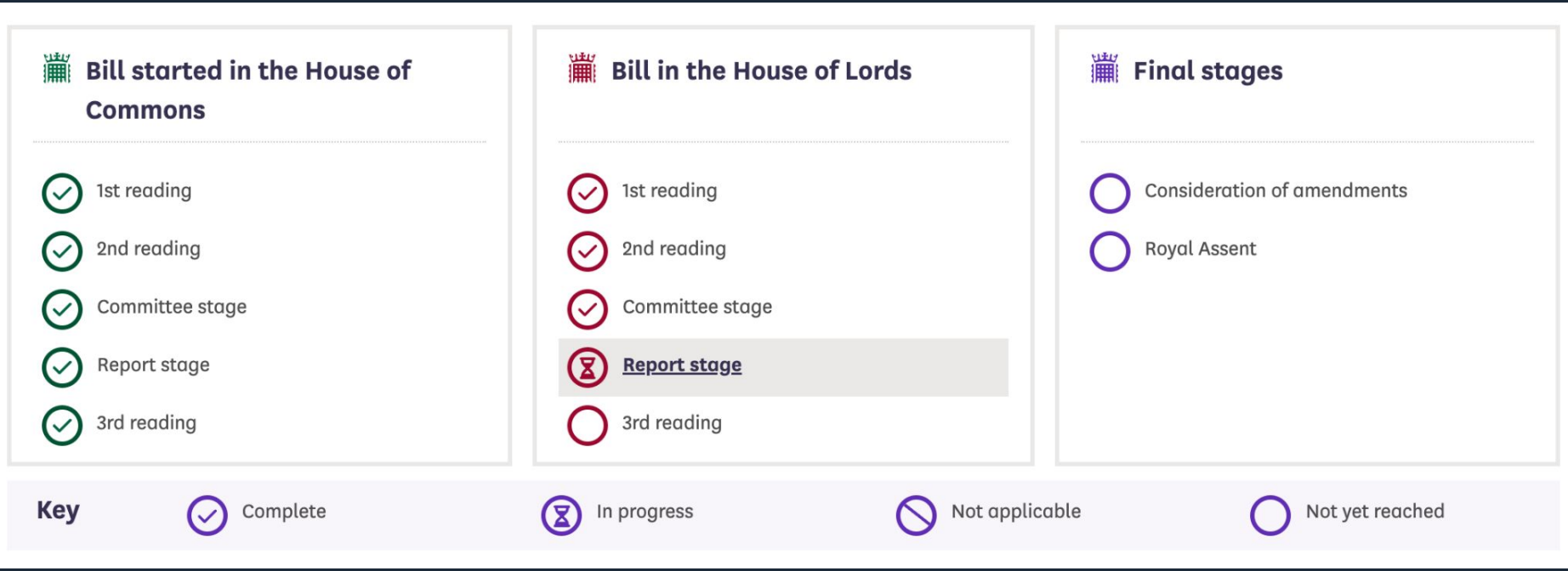
EU AI Act – High Risk

- Biometric
- Critical infrastructure
- Education
- Employment
- Credit scoring (except fraud)
- Law enforcement (general)
- Border control
- Justice
- … social media recommendation algorithms
- … “foundation models”

EU AI Act – Requirements

- Risk management processes and procedures
- Governance process
- Effective machine learning lifecycle management
- Robust deployment, logging, monitoring and auditing capabilities
- Right to explanation
- Reporting to authorities
- Comply to information and audit requests

What's Next?



winder.AI

winder.AI

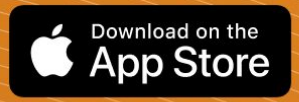
phil@winder.ai

https://Winder.AI

GOTO
Guide

Remember to **rate this session**

THANK YOU!



#GOTOams

Further Information

- Online Safety Bill – https://bills.parliament.uk/bills/3137
- Guide to the Online Safety Bill – https://www.gov.uk/guidance/a-guide-to-the-online-safety-bill

Title and Content

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore.

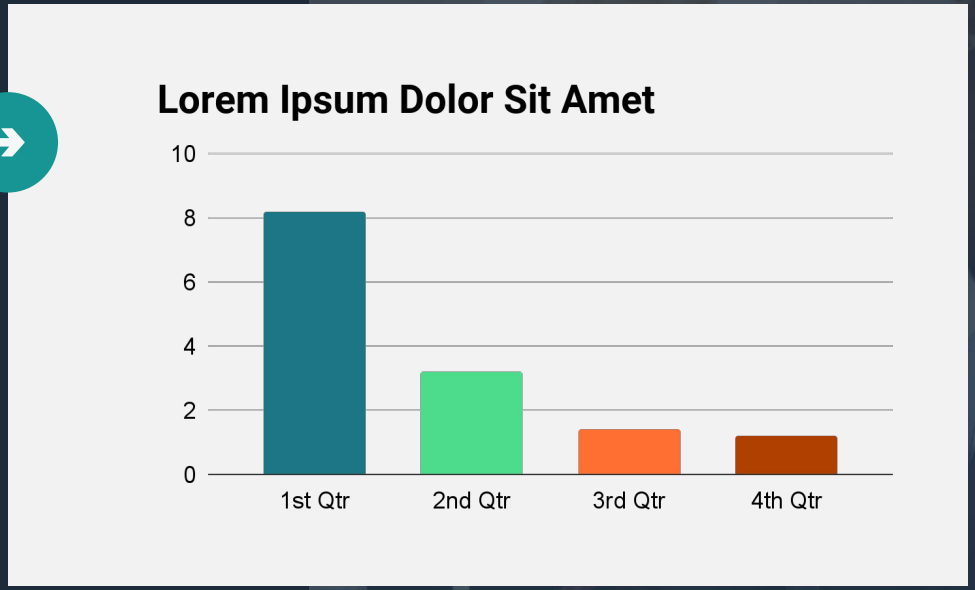
01

02

winder.AI

Title and Two Content

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor



winder.AI

Lorem ipsum dolor sit amet,
consectetur adipiscing elit

Title Slide

Lorem Ipsum Dolor Sit Amet

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Lorem Ipsum Dolor Sit Amet

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Lorem Ipsum Dolor Sit Amet

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Lorem Ipsum Dolor Sit Amet

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Picture Slide

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut.

- Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor.
- Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor.
- Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor.

winder.AI

This is Closing Slide