From: Media OeSC
Sent: Tuesday, 12 September 2023 7:04 AM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

s 22

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for

Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW


More like this


Less like this


Comment

From: Media OeSC
Sent: Tuesday, 12 September 2023 7:04 AM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

Follow Up Flag: Follow up
Flag Status: Completed

s 22

s 22

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)

- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW


More like this


Less like this


Comment

From: Media OeSC
Sent: Tuesday, 12 September 2023 7:04 AM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

s 22

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for

Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW


More like this


Less like this


Comment

From: Media OeSC
Sent: Tuesday, 12 September 2023 7:04 AM
To: s 47E(c), s 47F
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

s 22

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for

Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW


More like this


Less like this


Comment

From: s 47E(c), s 47F
Sent: Tuesday, 12 September 2023 7:04 AM
To: Julie Inman Grant s 47E(c), s 47F
Subject: Re: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

OFFICIAL

s 22

s 22

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

> You don't often get email from noreply@everythinginmoderation.co. Learn why this is important

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for

Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW


More like this


Less like this


Comment

From: s 47E(c), s 47F
Sent: Tuesday, 12 September 2023 6:33 AM
To: Julie Inman Grant; s 47E(c), s 47F
Subject: Re: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for

Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

From: Media OeSC
Sent: Monday, 11 September 2023 10:32 PM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for

Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW


More like this


Less like this


Comment

From: Media OeSC
Sent: Monday, 11 September 2023 10:32 PM
To: s 47E(c), s 47F
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

> You don't often get email from noreply@everythinginmoderation.co. Learn why this is important

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for

Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW


More like this


Less like this


Comment

From: Media OeSC
Sent: Monday, 11 September 2023 10:32 PM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for

Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW


More like this


Less like this


Comment

From: Media OeSC
Sent: Monday, 11 September 2023 10:32 PM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

> You don't often get email from noreply@everythinginmoderation.co. Learn why this is important

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for

Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW


More like this


Less like this


Comment

From: Julie Inman Grant
Sent: Monday, 11 September 2023 10:32 PM
To: s 47E(c), s 47F [redacted]
Subject: Re: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22 [redacted]

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for

Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW


More like this


Less like this


Comment

From: s 47E(c), s 47E
Sent: Monday, 11 September 2023 4:07 PM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

OFFICIAL

s 22

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

> You don't often get email from noreply@everythinginmoderation.co. Learn why this is important

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for

Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)

- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

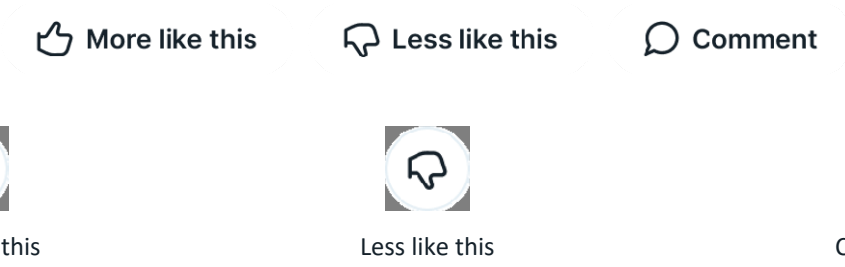
Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW



Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: s 47E(c), s 47E
Sent: Monday, 11 September 2023 9:09 AM
To: s 47E(c), s 47E
Subject: RE: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

OFFICIAL

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

> You don't often get email from noreply@everythinginmoderation.co. Learn why this is important

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for

Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)

- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

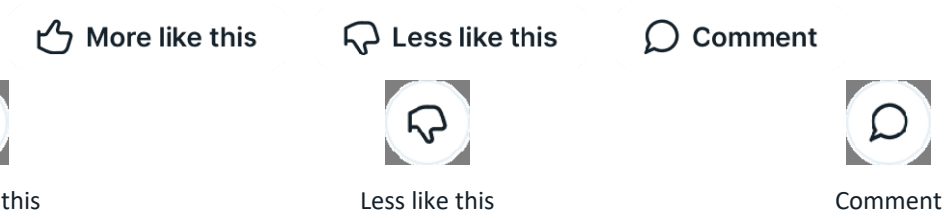
Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW



From: Media OeSC
Sent: Sunday, 10 September 2023 7:38 PM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

s 22

s 22

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Every week, industry experts and leaders read the weekly newsletter from start to end. Here are some nice things they've said:
- *"If you care about any of this and are not subscribing, you should!"* - Ashley
- *"He doesn't miss the little-covered but important stories"* - Kevin
- *"Every edition should be considered must-read content for those in the trust & safety space"* - Justin

Becoming a member helps me continue to curate the information and ideas you need to make the internet a safer, better place for everyone. Without its current members, EiM would have folded by now. So thanks to all members, past, current and future — BW

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)

- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is

forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.

- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: Media OeSC
Sent: Sunday, 10 September 2023 7:38 PM
To: s 47E(c), s 47F
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

s 22

s 22

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Every week, industry experts and leaders read the weekly newsletter from start to end. Here are some nice things they've said:
- *"If you care about any of this and are not subscribing, you should!"* - Ashley
- *"He doesn't miss the little-covered but important stories"* - Kevin
- *"Every edition should be considered must-read content for those in the trust & safety space"* - Justin

Becoming a member helps me continue to curate the information and ideas you need to make the internet a safer, better place for everyone. Without its current members, EiM would have folded by now. So thanks to all members, past, current and future — BW

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)

- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is

forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.

- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: Media OeSC
Sent: Sunday, 10 September 2023 7:38 PM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

s 22

s 22

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Every week, industry experts and leaders read the weekly newsletter from start to end. Here are some nice things they've said:
- *"If you care about any of this and are not subscribing, you should!"* - Ashley
- *"He doesn't miss the little-covered but important stories"* - Kevin
- *"Every edition should be considered must-read content for those in the trust & safety space"* - Justin

Becoming a member helps me continue to curate the information and ideas you need to make the internet a safer, better place for everyone. Without its current members, EiM would have folded by now. So thanks to all members, past, current and future — BW

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)

- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is

forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.

- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: Media OeSC
Sent: Sunday, 10 September 2023 7:38 PM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

s 22

s 22

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Every week, industry experts and leaders read the weekly newsletter from start to end. Here are some nice things they've said:
- *"If you care about any of this and are not subscribing, you should!"* - Ashley
- *"He doesn't miss the little-covered but important stories"* - Kevin
- *"Every edition should be considered must-read content for those in the trust & safety space"* - Justin

Becoming a member helps me continue to curate the information and ideas you need to make the internet a safer, better place for everyone. Without its current members, EiM would have folded by now. So thanks to all members, past, current and future — BW

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)

- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is

forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.

- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: s 47E(c), s 47F
Sent: Sunday, 10 September 2023 7:38 PM
To: Julie Inman Grant s 47E(c), s 47F
Subject: Re: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

OFFICIAL

s 22

s 22

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

- *"Every edition should be considered must-read content for those in the trust & safety space"* - Justin

Becoming a member helps me continue to curate the information and ideas you need to make the internet a safer, better place for everyone. Without its current members, EiM would have folded by now. So thanks to all members, past, current and future — BW

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)

- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: Media OeSC
Sent: Sunday, 10 September 2023 6:28 PM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

> You don't often get email from noreply@everythinginmoderation.co. Learn why this is important

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for

Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)

- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

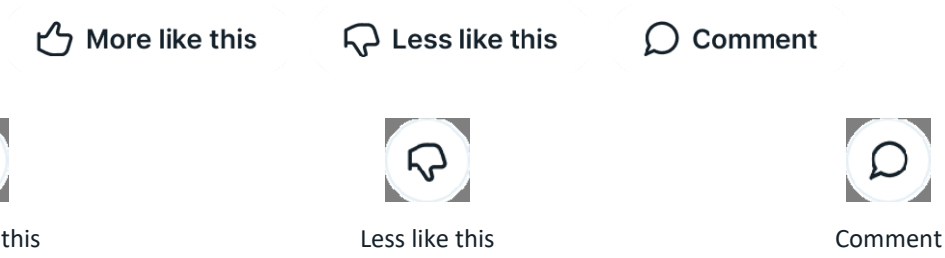
Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW



From: Media OeSC
Sent: Sunday, 10 September 2023 6:28 PM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

> You don't often get email from noreply@everythinginmoderation.co. Learn why this is important

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for

Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)

- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

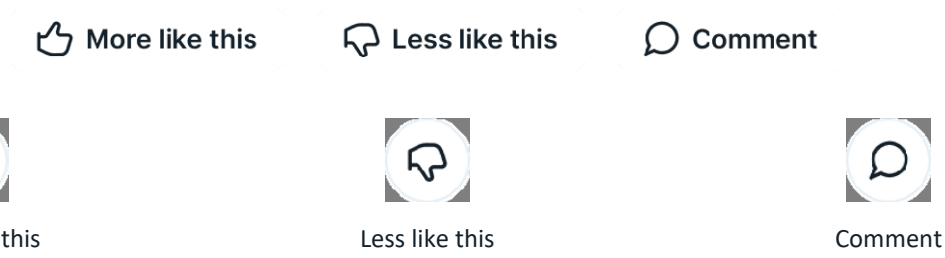
Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW



From: Media OeSC
Sent: Sunday, 10 September 2023 6:28 PM
To: s 47E(c), s 47F
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

> You don't often get email from noreply@everythinginmoderation.co. Learn why this is important

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for

Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)

- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

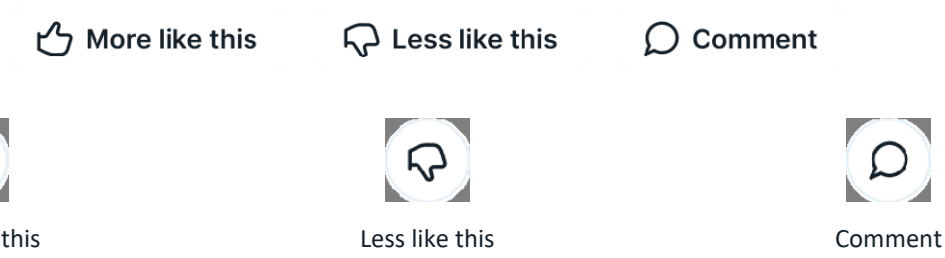
Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW



From: Media OeSC
Sent: Sunday, 10 September 2023 6:28 PM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

> You don't often get email from noreply@everythinginmoderation.co. Learn why this is important

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for

Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)

- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

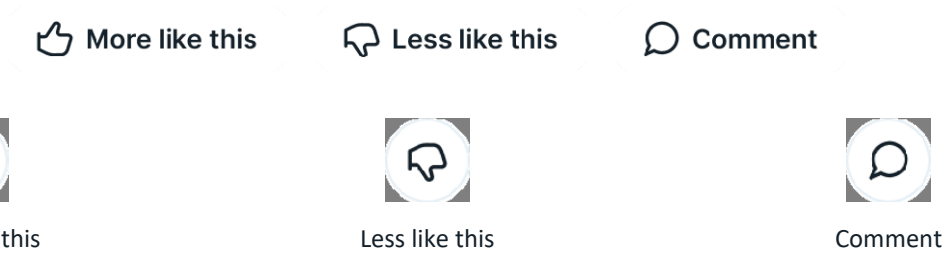
Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW



From: Julie Inman Grant
Sent: Sunday, 10 September 2023 6:28 PM
To: s 47E(c), s 47F [redacted]
Subject: Re: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22 [redacted]

s 22

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

> You don't often get email from noreply@everythinginmoderation.co. Learn why this is important

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in

the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing

market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety.

Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

More like this | Less like this | Comment

More like this

Less like this

Comment

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: s 47E(c), s 47E
Sent: Sunday, 10 September 2023 4:34 PM
To: s 47E(c), s 47E
Subject: RE: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

OFFICIAL

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

> You don't often get email from noreply@everythinginmoderation.co. Learn why this is important

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The*

Financial Times reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing

market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety.

Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

More like this

Less like this

Comment

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

Sent: Sunday, 10 September 2023 4:11 PM
To: s 47E(c), s 47E
Subject: RE: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

> You don't often get email from noreply@everythinginmoderation.co. Learn why this is important

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were

sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Become an EiM member
Everything in Moderation is your guide to understanding how content moderation is changing the world.

Every week, industry experts and leaders read the weekly newsletter from start to end. Here are some nice things they've said:
- *"If you care about any of this and are not subscribing, you should!"* - Ashley
- *"He doesn't miss the little-covered but important stories"* - Kevin
- *"Every edition should be considered must-read content for those in the trust & safety space"* - Justin

Becoming a member helps me continue to curate the information and ideas you need to make the internet a safer, better place for everyone. Without its current members, EiM would have folded by now. So thanks to all members, past, current and future — BW

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)

- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".

- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: Media OeSC
Sent: Sunday, 10 September 2023 8:23 AM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>

Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)

- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: Media OeSC
Sent: Sunday, 10 September 2023 8:23 AM
To: s 47E(c), s 47F
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>

Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)

- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: Media OeSC
Sent: Sunday, 10 September 2023 8:23 AM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>

Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)

- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: Media OeSC
Sent: Sunday, 10 September 2023 8:23 AM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

s 22

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>

Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)

- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: s 47E(c), s 47E
Sent: Sunday, 10 September 2023 8:23 AM
To: Julie Inman Grant;Media OeSC; s 47E(c), s 47F
Subject: RE: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list [SEC=OFFICIAL]

OFFICIAL

s 22

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

> You don't often get email from noreply@everythinginmoderation.co. Learn why this is important

The week in content moderation - edition #215

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for

Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)

- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

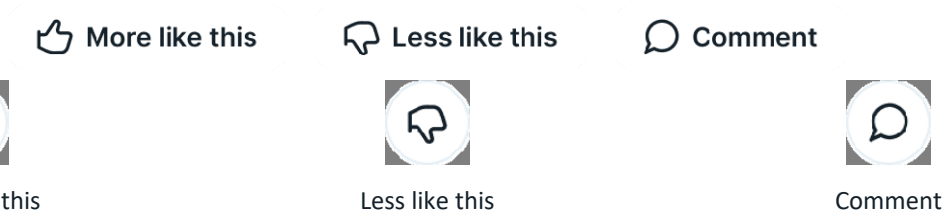
Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW



From: Media OeSC
Sent: Friday, 8 September 2023 10:45 PM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

s 22 [redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

EVERYTHING IN MODERATION*

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The*

Financial Times reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing

market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet

safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

More like this | Less like this | Comment

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: Media OeSC
Sent: Friday, 8 September 2023 10:45 PM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

EVERYTHING IN MODERATION*

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The*

Financial Times reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing

market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet

safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

More like this | Less like this | Comment

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: Media OeSC
Sent: Friday, 8 September 2023 10:45 PM
To: s 47E(c), s 47E
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

s 22

[redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

EVERYTHING IN MODERATION*

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The*

Financial Times reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing

market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet

safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

More like this | Less like this | Comment

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: Media OeSC
Sent: Friday, 8 September 2023 10:45 PM
To: s 47E(c), s 47F
Subject: FW: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

s 22

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

EVERYTHING IN MODERATION*

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The*

Financial Times reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing

market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet

safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

More like this | Less like this | Comment

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: Julie Inman Grant
Sent: Friday, 8 September 2023 10:45 PM
To: Media OeSC s 47E(c), s 47F [redacted]
Subject: Fwd: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

s 22 [redacted]

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, September 8, 2023 10:02 pm
To: Julie Inman Grant <Julie.InmanGrant@eSafety.gov.au>
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

EVERYTHING IN MODERATION*

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked into the long grass. However, others were

sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Become an EiM member
Everything in Moderation is your guide to understanding how content moderation is changing the world.

Every week, industry experts and leaders read the weekly newsletter from start to end. Here are some nice things they've said:
- *"If you care about any of this and are not subscribing, you should!"* - Ashley
- *"He doesn't miss the little-covered but important stories"* - Kevin
- *"Every edition should be considered must-read content for those in the trust & safety space"* - Justin

Becoming a member helps me continue to curate the information and ideas you need to make the internet a safer, better place for everyone. Without its current members, EiM would have folded by now. So thanks to all members, past, current and future — BW

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (w*arning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)

- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".

- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: noreply=everythinginmoderation.co@m.ghost.io on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, 8 September 2023 10:01 PM
To: s 47E(c), s 47E
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked

into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (*warning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around

the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

From: noreply=everythinginmoderation.co@m.ghost.io on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, 8 September 2023 10:01 PM
To: s 47E(c), s 47F
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

EVERYTHING IN MODERATION*

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked

into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (*warning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around

the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

From: noreply=everythinginmoderation.co@m.ghost.io on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, 8 September 2023 10:01 PM
To: s 47E(c), s 47E
Subject: UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

EVERYTHING IN MODERATION*

UK safety bill "not technically feasible", ActiveFence acquires rival and familiar faces on AI list

By Ben Whitelaw • 8 Sept 2023 •

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

The thread that ties together this week's two biggest safety stories is realisation; for the UK government, there's a growing realisation that its regulatory ambitions may go beyond what's technically possible. And for two trust and safety startups, it's a literal realisation, a coming together via acquisition. Read on for more.

What is your view on these stories and today's edition? You hit the thumbs at the end of the newsletter, drop me an email or share today's edition via social media with your hottest take. I'm particularly interested in hearing from the newest EiM subscribers from Cinder, the Council for Foreign Relations, Spotify, Holistic AI, Mozilla Foundation, Linklaters, Tech Against Terrorism, DCMS and a host of others.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

The UK government has seemingly conceded that the technology needed to scan encrypted messages and enforce the **Online Safety Bill** is "not technically feasible" at this point in time. *The Financial Times* reported comments from a junior minister during the third reading of the bill in the House of Lords, which came just weeks after secure messaging services threatened to pull out of the UK if the bill — described earlier this week by *Wired* as "a bad law" — was passed.

The story got a mixed reaction; privacy advocates saw it as win in the bill's long-running saga and possibly a sign that the legislation would be kicked

into the long grass. However, others were sceptical of both the intended meaning of the comments and their supposed impact on the final legislation.

Pat Walshe (aka Privacy Matters) has helpfully compared government statements from July and this week in a thread and the exercise goes to show why both sides are claiming victory. While we figure out what was actually said and what it means, please enjoy this response from an FT commenter:

"Technology exists to open letters and reseal them. Next the Government will propose to open all letters, scan them using OCR software, and search for "harmful" content."

Also in this section...

- Examining Australia's bid to curb online disinformation (*Australian Strategic Policy Institute*)
- European digital diplomacy, the way forward (*Telefonica*)
- New Research: Hate Speech Hurts Social Media Sites, Brands, and the Digital Economy (*Project DisCo*)

Products

Features, functionality and technology shaping online speech

Not quite on the level of the Online Safety Bill news but still a notable story: **ActiveFence** this week announced that it has acquired AI safety model company **Spectrum Labs** in what is "the largest M&A within the Trust & Safety industry", according to a press release. The move will allow ActiveFence's customers — which now include Riot Games, Grindr and Match Group — "to identify and mitigate potential online harms on an even greater scale, and faster".

The move, which will also see ActiveFence take on Spectrum's #TSCollective of 850+ members, marks what *TechCrunch*'s Ingrid Lunden calls "a classic development in what is a maturing market". I wouldn't be surprised if more of the smaller, emerging players in the space found new homes in the next 12-18 months.

Also in this section...

- X's Community Notes feature will now include videos (*The Verge*)
- Negative feedback could moderate social media extremes (*University of Michigan*)

Platforms

Social networks and the application of content guidelines

The outsourcing firm that moderated content for **Meta** until March this year has been accused of protecting an employee that allegedly raped two colleagues in Kenya and then sacked of his victims. Two former workers at **Sama**, formerly Samasource, spoke to OpenDemocracy (*warning: includes descriptions of sexual assault and rape*) about the incident and criticised the company counseller's attempt to "repeatedly tried to persuade me [the victim] to forgive him". They are now part of one of a number of ongoing lawsuits against Meta going through the courts in Kenya (EiM #153).

The allegations come in the same week that the country's trade cabinet secretary Moses Kuria reportedly revealed that Sama will hire more than 2000 workers to label images and videos for machine learning algorithms. Let's hope the standard of care for workers has improved.

Bumble has released a fresh version of its Community Guidelines as part of a continued effort to "be a space to make kind connections in a safe, inclusive, and respectful way". Interestingly, the dating app has categorised no shows — that is, agreeing to meet up but never showing — as Bullying and Abusive Conduct and has also banned the use of automation and scripting to artificially influence "connections, matching, conversations or engagement".

Also in this section...

- Reddit faces content quality concerns after its Great Mod Purge (*ArsTechnica*)
- San Jose mayor wants Meta, Snapchat and TikTok to shut down sideshow content (*Mercury News*)
- The endless battle to banish the world's most notorious stalker website (*The Washington Post*)
- Meta's Mark Zuckerberg faces battles in Irish court (*The Times*)
- Child Protection Professionals Censored on Wikipedia (*Jeremy Malcolm*)
- Inside Musk's Twitter Transformation: Impulsive Decisions, Favors for Friends (*The Wall Street Journal*)
- How We're Making Alcohol Deliveries Even Safer (*DoorDash*)

People

Those impacting the future of online safety and moderation

While working at Sama, **Richard Mathenge** helped train the AI model that powered OpenAI's ChatGPT, which exploded earlier this year. Then, in May this year, he helped form the African Content Moderators Union (EiM #199), the first collective of its kind.

Now, along with the likes of Rumman Chowdhury, Timnit Gebru and even Grimes, Mathenge is part of TIME's list of the 100 most influential people in AI. The newly published list brings together "industry leaders at the forefront of the AI boom, individuals outside these companies who are grappling with profound ethical questions around the uses of AI, and the innovators around

the world who are trying to use AI to address social challenges". How great it is to see Richard being included such an illustrious list.

He's not the only one in the list from the trust and safety space, either. A number of the founders of Anthropic, the AI safety-research lab, are there (also featured in *Forbes* this week) as well as the executive director for the Center for AI Safety.

Posts of note

Handpicked posts that caught my eye this week

- "it is very funny that community notes, which began as a 'trust & safety feature', has slowly just become another avenue for quote-tweet dunks" - **Will Partin**, who works at YouTube's hate speech team, notes that Elon has ruined what could've been "a good thing".
- "It is deeply disturbing that antisemitic, racist & hateful words posted here do not break safety policies on this platform." - When a national memorial with 1.5 million followers is forced to tweet to post at the CEO to get attention about hate speech, you know something isn't right.
- "Advertisers *really* don't want to advertise next to hate speech." - Progress Chamber founder and CEO **Adam Kovacevich** shares a new study on the impact of Florida and Texas anti-moderation laws.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 1900+ EiM subscribers.

OpenAI is looking for a UK Policy and Partnerships Lead to be part of its Global Affairs team to lead it's UK public policy engagement.

The role, which reports to the Head of European Policy and Partnerships, involves managing relationships with key stakeholders, shaping strategic initiatives and representing the company in public and private forums.

The right person will have deep expertise in technology policy, excellent communication and interpersonal skills and a track record of working with cross-functional teams. No salary on the LinkedIn ad but, unsurprisingly, there's already a ton of interest.

Every week, I read and summarise the need-to-know articles, analyses and research about content moderation to give you a broader, more global perspective on online speech and internet safety. Support my work and the newsletter by becoming a member for less than $2 a week or, if you can't do that, forwarding this edition to someone like you who cares about the future of the web. Thanks for reading - BW

From: s 47E(c), s 47F
Sent: Tuesday, 15 August 2023 9:09 AM
To: s 47E(c), s 47F
Subject: FW: 'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk [SEC=OFFICIAL]

Follow Up Flag: Follow up
Flag Status: Flagged

OFFICIAL

s 22

From: noreply=everythinginmoderation.co@m.ghost.io <noreply=everythinginmoderation.co@m.ghost.io> **On Behalf Of** Ben from Everything in Moderation*
Sent: Friday, 11 August 2023 10:01 PM
To: s 47E(c), s 47F
Subject: 'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

The week in content moderation - edition #211

'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

This edition comes after a two-week break due to a holiday and then an unexpected hospital trip last Friday. Everyone and everything is fine but I'm sorry to those of you who were expecting their weekly newsletter to arrive and who were surprised when it didn't.

Today's newsletter shows, once again, the very tangible impact of incoming online speech regulation, both in the ways that it is shaping platform features as well as how it will shape future legislation.

Before we get into the big stories, I want to welcome a host of new subscribers since the last EiM from Google, Checkstep, Uber, TrustLab, Electronic Arts, Feeld, Mailchimp, Bumble, Shopify, Etsy and others. Your feedback is key to the newsletter hitting the mark each week. Email me or hit the thumbs at the end of today's edition.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

A new report has found that the **terms and conditions** of video-sharing platforms are "lengthy, impenetrable and, in some cases, inconsistent" and "risk leaving users and moderators in the dark." UK regulator Ofcom analysed OnlyFans, Twitch, Snapchat, TikTok, Brand New Tube and BitChute's terms and

found each of them was "difficult to read" and would take up to an hour to get through. They also noted that few provide detailed information about content that violates the terms or penalties for breaking the rules.

Helping users better understand platforms' terms of service was one of the key themes at last month's Trust and Safety hackathon attended by dozens of trust and safety workers ahead of TrustCon (I was there too). If I was one of the video-sharing platforms mentioned in the report, I'd be taking a closer look.

New research looking at Facebook's interventions during the 2020 US election "highlight the fallibility of corporate collaboration and the need for government **transparency** mandates", according to an analysis by Justin Hendrix, CEO and editor of *Tech Policy Press*, and Paul Barrett, deputy director at the Center for Business and Human Rights at NYU Stern. Looking at the papers published in Science and Nature as well as media coverage, the pair raised questions about whether "Meta may be keen to show that its model in this collaboration can inform such rules" as the Digital Services Act and proposed legislation in the US. My read of the week.

Also in this section...

- The Internet Speech Case That the Supreme Court Can't Dodge (*Wired*)
- Coming regulations mean game developers must be proactive on safety and trust (*Venturebeat*)
- The Lawfare Podcast: Can We Build a Trustworthy Future Web? (*Lawfare*)
- France: Proposed internet bill threatens online speech (*Article19*)
- EU tech envoy: 'The winds have changed' on regulating Silicon Valley (*Politico*)

Products

Features, functionality and technology shaping online speech

TikTok users will soon be able to opt out of a **personalised feed** as part of the platform's effort to comply with the Digital Services Act. According to a blog post, users can choose to be served popular content from where they live and

to have chronological content from Following and Friends. It will also update its reporting flows to allow users to flag content they believe is illegal, another DSA requirement, as well as "provide our community in Europe with information about a broader range of content moderation decisions"; however, details about how this works in practice are scarce.

Platforms

Social networks and the application of content guidelines

Former moderators working for Sama on **OpenAI**'s ChatGPT output have spoken out about their working conditions and the efforts of their job reviewing and testing large language models. This report from *The Guardian* quotes two moderators and explains the fallout of reading violent and sexualised text snippets all day: group calls to swap horror stories, hobbies falling by the wayside and couples splitting up. Back in July, four moderators filed a petition to Kenya's National Assembly calling for an investigation into the working conditions of outsourced workers on AI content moderation.

Also in this section...

- Twitch expands its ban on gambling livestreams (*NBC News*)

- Musk says X to fund legal bills if users 'unfairly treated' over posts (*Al Jazeera*)
- Accenture: tech's silent partner cuts deep in Ireland (*The Times*)
- X tries to win back advertisers with brand safety promises (*The Register*)

People

Those impacting the future of online safety and moderation

The Center for Countering Hate and its research output has attracted significant media coverage since its founding in 2018. It has shone a light on platforms' failures to stop the spread of hate (EiM #124) so it was a matter of time before Elon Musk took aim at the British-based NGO.

Twitter/X's CEO called the organisation "evil" but also went after its founder and CEO **Imran Ahmed**, who he called a "rat". Ahmed is a former senior political advisor for the Labour party and, before that worked for Merrill Lunch and as a strategy consultant for the best part of a decade.

He writes in *The Guardian* this week that "attacking CCDH will not remove the neo-Nazis, white supremacists and disinformation superspreaders he has allowed back". And he warns that "If we allow ourselves to be strongarmed by Musk and X, it will give the green light to every other social media behemoth to do the same to anyone who dares to speak truth to power.".

Posts of note

Handpicked posts that caught my eye this week

- "What about the Trust and Safety Council we've asked her to reinstate?!" - **Global Project Against Hate and Extremism** wonders whether X's reformed 'client council' is a sign of things to come (narrator: it's not)
- "These models will cause big Online Safety Bill headaches." - **Matthew Feeney**, head of technology and innovation at CPS Think Tank, on new research that shows AI models have political biases.

- "AI generated child sex images: a "nightmare" "worst case scenario" for trust and safety? I argue otherwise" - former Protasia Foundation exec director and trust and safety consultant **Jeremy Malcolm** with an interesting thread.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 2000+ EiM subscribers.

Roblox is looking for a Product Marketing Lead, Trust and Safety.

The role is a partner for its Safety Product Management teams —who are responsible for Roblox's policies and scaled operations— and will be responsible for positioning, messaging, content creation, channel planning and measurement for the Safety group's product and policy launches.

It's an interesting role as communications becomes recognised as a strategic trust & safety tool (c/o Mike Masnick over at *Techdirt*). Whoever is successful will have 7+ years of product marketing experience with a leading global technology platform, with experience in trust and safety, privacy, or policy preferred.

The role is based at Roblox HQ in San Mateo, California with a not insignificant (and oddly specific) salary of $212,530—260,790.

Enjoy today's edition ? If so...

1. ***Hit the thumbs** or send a quick reply to share feedback or just say hello*
2. ***Forward to a friend** or colleague (PS they can sign up here)*
3. ***Become an EiM member** and get access to original analysis and the whole EiM archive for less than the price of a coffee a week*

More like this

Less like this

Comment

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: noreply=everythinginmoderation.co@m.ghost.io on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, 11 August 2023 10:01 PM
To: s 47E(c), s 47E
Subject: 'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

EVERYTHING IN MODERATION*

'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

By Ben Whitelaw • 11 Aug 2023 •

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

This edition comes after a two-week break due to a holiday and then an unexpected hospital trip last Friday. Everyone and everything is fine but I'm sorry to those of you who were expecting their weekly newsletter to arrive and who were surprised when it didn't.

Today's newsletter shows, once again, the very tangible impact of incoming online speech regulation, both in the ways that it is shaping platform features as well as how it will shape future legislation.

Before we get into the big stories, I want to welcome a host of new subscribers since the last EiM from Google, Checkstep, Uber, TrustLab, Electronic Arts, Feeld, Mailchimp, Bumble, Shopify, Etsy and others. Your feedback is key to the newsletter hitting the mark each week. Email me or hit the thumbs at the end of today's edition.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

A new report has found that the **terms and conditions** of video-sharing platforms are "lengthy, impenetrable and, in some cases, inconsistent" and "risk leaving users and moderators in the dark." UK regulator Ofcom analysed OnlyFans, Twitch, Snapchat, TikTok, Brand New Tube and BitChute's terms and found each of them was "difficult to read" and would take up to an hour to get through. They also noted that few provide detailed information about content that violates the terms or penalties for breaking the rules.

Helping users better understand platforms' terms of service was one of the key themes at last month's Trust and Safety hackathon attended by dozens of

trust and safety workers ahead of TrustCon (I was there too). If I was one of the video-sharing platforms mentioned in the report, I'd be taking a closer look.

New research looking at Facebook's interventions during the 2020 US election "highlight the fallibility of corporate collaboration and the need for government **transparency** mandates", according to an analysis by Justin Hendrix, CEO and editor of *Tech Policy Press*, and Paul Barrett, deputy director at the Center for Business and Human Rights at NYU Stern. Looking at the papers published in Science and Nature as well as media coverage, the pair raised questions about whether "Meta may be keen to show that its model in this collaboration can inform such rules" as the Digital Services Act and proposed legislation in the US. My read of the week.

Also in this section...

- The Internet Speech Case That the Supreme Court Can't Dodge (*Wired*)
- Coming regulations mean game developers must be proactive on safety and trust (*Venturebeat*)
- The Lawfare Podcast: Can We Build a Trustworthy Future Web? (*Lawfare*)
- France: Proposed internet bill threatens online speech (*Article19*)
- EU tech envoy: 'The winds have changed' on regulating Silicon Valley (*Politico*)

Products

Features, functionality and technology shaping online speech

TikTok users will soon be able to opt out of a **personalised feed** as part of the platform's effort to comply with the Digital Services Act. According to a blog post, users can choose to be served popular content from where they live and to have chronological content from Following and Friends. It will also update its reporting flows to allow users to flag content they believe is illegal, another DSA requirement, as well as "provide our community in Europe with information about a broader range of content moderation decisions"; however, details about how this works in practice are scarce.

Become an EiM member
Everything in Moderation is your guide to understanding how content moderation is changing the world.

Every week, industry experts and leaders read the weekly newsletter from start to end. Here are some nice things they've said:
- "*If you care about any of this and are not subscribing, you should!*" - Ashley
- "*He doesn't miss the little-covered but important stories*" - Kevin
- "*Every edition should be considered must-read content for those in the trust & safety space*" - Justin

Becoming a member helps me continue to curate the information and ideas you need to make the internet a safer, better place for everyone. Without its current members, EiM would have folded by now. So thanks to all members, past, current and future — BW

Platforms

Social networks and the application of content guidelines

Former moderators working for Sama on **OpenAI**'s ChatGPT output have spoken out about their working conditions and the efforts of their job reviewing and testing large language models. This report from *The Guardian* quotes two moderators and explains the fallout of reading violent and sexualised text snippets all day: group calls to swap horror stories, hobbies falling by the wayside and couples splitting up. Back in July, four moderators filed a petition to Kenya's National Assembly calling for an investigation into the working conditions of outsourced workers on AI content moderation.

Also in this section...

- Twitch expands its ban on gambling livestreams (*NBC News*)
- Musk says X to fund legal bills if users 'unfairly treated' over posts (*Al Jazeera*)
- Accenture: tech's silent partner cuts deep in Ireland (*The Times*)
- X tries to win back advertisers with brand safety promises (*The Register*)

People

Those impacting the future of online safety and moderation

The Center for Countering Hate and its research output has attracted significant media coverage since its founding in 2018. It has shone a light on platforms' failures to stop the spread of hate (EiM #124) so it was a matter of time before Elon Musk took aim at the British-based NGO.

Twitter/X's CEO called the organisation "evil" but also went after its founder and CEO **Imran Ahmed**, who he called a "rat". Ahmed is a former senior political advisor for the Labour party and, before that worked for Merrill Lunch and as a strategy consultant for the best part of a decade.

He writes in *The Guardian* this week that "attacking CCDH will not remove the neo-Nazis, white supremacists and disinformation superspreaders he has allowed back". And he warns that "If we allow ourselves to be strongarmed by Musk and X, it will give the green light to every other social media behemoth to do the same to anyone who dares to speak truth to power.".

Posts of note

Handpicked posts that caught my eye this week

- "What about the Trust and Safety Council we've asked her to reinstate?!" - **Global Project Against Hate and Extremism** wonders whether X's reformed 'client council' is a sign of things to come (narrator: it's not)
- "These models will cause big Online Safety Bill headaches." - **Matthew Feeney**, head of technology and innovation at CPS Think Tank, on new research that shows AI models have political biases.
- "AI generated child sex images: a "nightmare" "worst case scenario" for trust and safety? I argue otherwise" - former Protasia Foundation exec director and trust and safety consultant **Jeremy Malcolm** with an interesting thread.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 2000+ EiM subscribers.

Roblox is looking for a Product Marketing Lead, Trust and Safety.

The role is a partner for its Safety Product Management teams —who are responsible for Roblox's policies and scaled operations— and will be responsible for positioning, messaging, content creation, channel planning and measurement for the Safety group's product and policy launches.

It's an interesting role as communications becomes recognised as a strategic trust & safety tool (c/o Mike Masnick over at *Techdirt*). Whoever is successful will have 7+ years of product marketing experience with a leading global technology platform, with experience in trust and safety, privacy, or policy preferred.

The role is based at Roblox HQ in San Mateo, California with a not insignificant (and oddly specific) salary of $212,530—260,790.

Enjoy today's edition ? If so...

1. ***Hit the thumbs*** *or send a quick reply to share feedback or just say hello*
2. ***Forward to a friend*** *or colleague (PS they can sign up here)*
3. ***Become an EiM member*** *and get access to original analysis and the whole EiM archive for less than the price of a coffee a week*

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: noreply=everythinginmoderation.co@m.ghost.io on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, 11 August 2023 10:01 PM
To: s 47E(c), s 47E
Subject: 'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

By Ben Whitelaw • 11 Aug 2023 •

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

This edition comes after a two-week break due to a holiday and then an unexpected hospital trip last Friday. Everyone and everything is fine but I'm sorry to those of you who were expecting their weekly newsletter to arrive and who were surprised when it didn't.

Today's newsletter shows, once again, the very tangible impact of incoming online speech regulation, both in the ways that it is shaping platform features as well as how it will shape future legislation.

Before we get into the big stories, I want to welcome a host of new subscribers since the last EiM from Google, Checkstep, Uber, TrustLab, Electronic Arts, Feeld, Mailchimp, Bumble, Shopify, Etsy and others. Your feedback is key to the newsletter hitting the mark each week. Email me or hit the thumbs at the end of today's edition.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

A new report has found that the **terms and conditions** of video-sharing platforms are "lengthy, impenetrable and, in some cases, inconsistent" and "risk leaving users and moderators in the dark." UK regulator Ofcom analysed OnlyFans, Twitch, Snapchat, TikTok, Brand New Tube and BitChute's terms and found each of them was "difficult to read" and would take up to an hour to get through. They also noted that few provide detailed information about content that violates the terms or penalties for breaking the rules.

Helping users better understand platforms' terms of service was one of the key themes at last month's Trust and Safety hackathon attended by dozens of

trust and safety workers ahead of TrustCon (I was there too). If I was one of the video-sharing platforms mentioned in the report, I'd be taking a closer look.

New research looking at Facebook's interventions during the 2020 US election "highlight the fallibility of corporate collaboration and the need for government **transparency** mandates", according to an analysis by Justin Hendrix, CEO and editor of *Tech Policy Press*, and Paul Barrett, deputy director at the Center for Business and Human Rights at NYU Stern. Looking at the papers published in Science and Nature as well as media coverage, the pair raised questions about whether "Meta may be keen to show that its model in this collaboration can inform such rules" as the Digital Services Act and proposed legislation in the US. My read of the week.

Also in this section...

- The Internet Speech Case That the Supreme Court Can't Dodge (*Wired*)
- Coming regulations mean game developers must be proactive on safety and trust (*Venturebeat*)
- The Lawfare Podcast: Can We Build a Trustworthy Future Web? (*Lawfare*)
- France: Proposed internet bill threatens online speech (*Article19*)
- EU tech envoy: 'The winds have changed' on regulating Silicon Valley (*Politico*)

Products

Features, functionality and technology shaping online speech

TikTok users will soon be able to opt out of a **personalised feed** as part of the platform's effort to comply with the Digital Services Act. According to a blog post, users can choose to be served popular content from where they live and to have chronological content from Following and Friends. It will also update its reporting flows to allow users to flag content they believe is illegal, another DSA requirement, as well as "provide our community in Europe with information about a broader range of content moderation decisions"; however, details about how this works in practice are scarce.

Become an EiM member
Everything in Moderation is your guide to understanding how content moderation is changing the world.

Every week, industry experts and leaders read the weekly newsletter from start to end. Here are some nice things they've said:
- "*If you care about any of this and are not subscribing, you should!*" - Ashley
- "*He doesn't miss the little-covered but important stories*" - Kevin
- "*Every edition should be considered must-read content for those in the trust & safety space*" - Justin

Becoming a member helps me continue to curate the information and ideas you need to make the internet a safer, better place for everyone. Without its current members, EiM would have folded by now. So thanks to all members, past, current and future — BW

Platforms

Social networks and the application of content guidelines

Former moderators working for Sama on **OpenAI**'s ChatGPT output have spoken out about their working conditions and the efforts of their job reviewing and testing large language models. This report from *The Guardian* quotes two moderators and explains the fallout of reading violent and sexualised text snippets all day: group calls to swap horror stories, hobbies falling by the wayside and couples splitting up. Back in July, four moderators filed a petition to Kenya's National Assembly calling for an investigation into the working conditions of outsourced workers on AI content moderation.

Also in this section...

- Twitch expands its ban on gambling livestreams (*NBC News*)
- Musk says X to fund legal bills if users 'unfairly treated' over posts (*Al Jazeera*)
- Accenture: tech's silent partner cuts deep in Ireland (*The Times*)
- X tries to win back advertisers with brand safety promises (*The Register*)

People

Those impacting the future of online safety and moderation

The Center for Countering Hate and its research output has attracted significant media coverage since its founding in 2018. It has shone a light on platforms' failures to stop the spread of hate (EiM #124) so it was a matter of time before Elon Musk took aim at the British-based NGO.

Twitter/X's CEO called the organisation "evil" but also went after its founder and CEO **Imran Ahmed**, who he called a "rat". Ahmed is a former senior political advisor for the Labour party and, before that worked for Merrill Lunch and as a strategy consultant for the best part of a decade.

He writes in *The Guardian* this week that "attacking CCDH will not remove the neo-Nazis, white supremacists and disinformation superspreaders he has allowed back". And he warns that "If we allow ourselves to be strongarmed by Musk and X, it will give the green light to every other social media behemoth to do the same to anyone who dares to speak truth to power.".

Posts of note

Handpicked posts that caught my eye this week

- "What about the Trust and Safety Council we've asked her to reinstate?!" - **Global Project Against Hate and Extremism** wonders whether X's reformed 'client council' is a sign of things to come (narrator: it's not)
- "These models will cause big Online Safety Bill headaches." - **Matthew Feeney**, head of technology and innovation at CPS Think Tank, on new research that shows AI models have political biases.
- "AI generated child sex images: a "nightmare" "worst case scenario" for trust and safety? I argue otherwise" - former Protasia Foundation exec director and trust and safety consultant **Jeremy Malcolm** with an interesting thread.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 2000+ EiM subscribers.

Roblox is looking for a Product Marketing Lead, Trust and Safety.

The role is a partner for its Safety Product Management teams —who are responsible for Roblox's policies and scaled operations— and will be responsible for positioning, messaging, content creation, channel planning and measurement for the Safety group's product and policy launches.

It's an interesting role as communications becomes recognised as a strategic trust & safety tool (c/o Mike Masnick over at *Techdirt*). Whoever is successful will have 7+ years of product marketing experience with a leading global technology platform, with experience in trust and safety, privacy, or policy preferred.

The role is based at Roblox HQ in San Mateo, California with a not insignificant (and oddly specific) salary of $212,530—260,790.

Enjoy today's edition ? If so...

1. ***Hit the thumbs*** *or send a quick reply to share feedback or just say hello*
2. ***Forward to a friend*** *or colleague (PS they can sign up here)*
3. ***Become an EiM member*** *and get access to original analysis and the whole EiM archive for less than the price of a coffee a week*

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: noreply=everythinginmoderation.co@m.ghost.io on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, 11 August 2023 10:01 PM
To: Julie Inman Grant
Subject: 'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

By Ben Whitelaw • 11 Aug 2023 •

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

This edition comes after a two-week break due to a holiday and then an unexpected hospital trip last Friday. Everyone and everything is fine but I'm sorry to those of you who were expecting their weekly newsletter to arrive and who were surprised when it didn't.

Today's newsletter shows, once again, the very tangible impact of incoming online speech regulation, both in the ways that it is shaping platform features as well as how it will shape future legislation.

Before we get into the big stories, I want to welcome a host of new subscribers since the last EiM from Google, Checkstep, Uber, TrustLab, Electronic Arts, Feeld, Mailchimp, Bumble, Shopify, Etsy and others. Your feedback is key to the newsletter hitting the mark each week. Email me or hit the thumbs at the end of today's edition.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

A new report has found that the **terms and conditions** of video-sharing platforms are "lengthy, impenetrable and, in some cases, inconsistent" and "risk leaving users and moderators in the dark." UK regulator Ofcom analysed OnlyFans, Twitch, Snapchat, TikTok, Brand New Tube and BitChute's terms and found each of them was "difficult to read" and would take up to an hour to get through. They also noted that few provide detailed information about content that violates the terms or penalties for breaking the rules.

Helping users better understand platforms' terms of service was one of the key themes at last month's Trust and Safety hackathon attended by dozens of

trust and safety workers ahead of TrustCon (I was there too). If I was one of the video-sharing platforms mentioned in the report, I'd be taking a closer look.

New research looking at Facebook's interventions during the 2020 US election "highlight the fallibility of corporate collaboration and the need for government **transparency** mandates", according to an analysis by Justin Hendrix, CEO and editor of *Tech Policy Press*, and Paul Barrett, deputy director at the Center for Business and Human Rights at NYU Stern. Looking at the papers published in Science and Nature as well as media coverage, the pair raised questions about whether "Meta may be keen to show that its model in this collaboration can inform such rules" as the Digital Services Act and proposed legislation in the US. My read of the week.

Also in this section...

- The Internet Speech Case That the Supreme Court Can't Dodge (*Wired*)
- Coming regulations mean game developers must be proactive on safety and trust (*Venturebeat*)
- The Lawfare Podcast: Can We Build a Trustworthy Future Web? (*Lawfare*)
- France: Proposed internet bill threatens online speech (*Article19*)
- EU tech envoy: 'The winds have changed' on regulating Silicon Valley (*Politico*)

Products

Features, functionality and technology shaping online speech

TikTok users will soon be able to opt out of a **personalised feed** as part of the platform's effort to comply with the Digital Services Act. According to a blog post, users can choose to be served popular content from where they live and to have chronological content from Following and Friends. It will also update its reporting flows to allow users to flag content they believe is illegal, another DSA requirement, as well as "provide our community in Europe with information about a broader range of content moderation decisions"; however, details about how this works in practice are scarce.

Become an EiM member
Everything in Moderation is your guide to understanding how content moderation is changing the world.

Every week, industry experts and leaders read the weekly newsletter from start to end. Here are some nice things they've said:
- "*If you care about any of this and are not subscribing, you should!*" - Ashley
- "*He doesn't miss the little-covered but important stories*" - Kevin
- "*Every edition should be considered must-read content for those in the trust & safety space*" - Justin

Becoming a member helps me continue to curate the information and ideas you need to make the internet a safer, better place for everyone. Without its current members, EiM would have folded by now. So thanks to all members, past, current and future — BW

Platforms

Social networks and the application of content guidelines

Former moderators working for Sama on **OpenAI**'s ChatGPT output have spoken out about their working conditions and the efforts of their job reviewing and testing large language models. This report from *The Guardian* quotes two moderators and explains the fallout of reading violent and sexualised text snippets all day: group calls to swap horror stories, hobbies falling by the wayside and couples splitting up. Back in July, four moderators filed a petition to Kenya's National Assembly calling for an investigation into the working conditions of outsourced workers on AI content moderation.

Also in this section...

- Twitch expands its ban on gambling livestreams (*NBC News*)
- Musk says X to fund legal bills if users 'unfairly treated' over posts (*Al Jazeera*)
- Accenture: tech's silent partner cuts deep in Ireland (*The Times*)
- X tries to win back advertisers with brand safety promises (*The Register*)

People

Those impacting the future of online safety and moderation

The Center for Countering Hate and its research output has attracted significant media coverage since its founding in 2018. It has shone a light on platforms' failures to stop the spread of hate (EiM #124) so it was a matter of time before Elon Musk took aim at the British-based NGO.

Twitter/X's CEO called the organisation "evil" but also went after its founder and CEO **Imran Ahmed**, who he called a "rat". Ahmed is a former senior political advisor for the Labour party and, before that worked for Merrill Lunch and as a strategy consultant for the best part of a decade.

He writes in *The Guardian* this week that "attacking CCDH will not remove the neo-Nazis, white supremacists and disinformation superspreaders he has allowed back". And he warns that "If we allow ourselves to be strongarmed by Musk and X, it will give the green light to every other social media behemoth to do the same to anyone who dares to speak truth to power.".

Posts of note

Handpicked posts that caught my eye this week

- "What about the Trust and Safety Council we've asked her to reinstate?!" - **Global Project Against Hate and Extremism** wonders whether X's reformed 'client council' is a sign of things to come (narrator: it's not)
- "These models will cause big Online Safety Bill headaches." - **Matthew Feeney**, head of technology and innovation at CPS Think Tank, on new research that shows AI models have political biases.
- "AI generated child sex images: a "nightmare" "worst case scenario" for trust and safety? I argue otherwise" - former Protasia Foundation exec director and trust and safety consultant **Jeremy Malcolm** with an interesting thread.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 2000+ EiM subscribers.

Roblox is looking for a Product Marketing Lead, Trust and Safety.

The role is a partner for its Safety Product Management teams —who are responsible for Roblox's policies and scaled operations— and will be responsible for positioning, messaging, content creation, channel planning and measurement for the Safety group's product and policy launches.

It's an interesting role as communications becomes recognised as a strategic trust & safety tool (c/o Mike Masnick over at *Techdirt*). Whoever is successful will have 7+ years of product marketing experience with a leading global technology platform, with experience in trust and safety, privacy, or policy preferred.

The role is based at Roblox HQ in San Mateo, California with a not insignificant (and oddly specific) salary of $212,530—260,790.

Enjoy today's edition Julie? *If so...*

1. ***Hit the thumbs*** *or send a quick reply to share feedback or just say hello*
2. ***Forward to a friend*** *or colleague (PS they can sign up here)*
3. ***Become an EiM member*** *and get access to original analysis and the whole EiM archive for less than the price of a coffee a week*

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: noreply=everythinginmoderation.co@m.ghost.io on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, 11 August 2023 10:01 PM
To: s 47E(c), s 47F
Subject: 'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

By Ben Whitelaw • 11 Aug 2023 •

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

This edition comes after a two-week break due to a holiday and then an unexpected hospital trip last Friday. Everyone and everything is fine but I'm sorry to those of you who were expecting their weekly newsletter to arrive and who were surprised when it didn't.

Today's newsletter shows, once again, the very tangible impact of incoming online speech regulation, both in the ways that it is shaping platform features as well as how it will shape future legislation.

Before we get into the big stories, I want to welcome a host of new subscribers since the last EiM from Google, Checkstep, Uber, TrustLab, Electronic Arts, Feeld, Mailchimp, Bumble, Shopify, Etsy and others. Your feedback is key to the newsletter hitting the mark each week. Email me or hit the thumbs at the end of today's edition.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

A new report has found that the **terms and conditions** of video-sharing platforms are "lengthy, impenetrable and, in some cases, inconsistent" and "risk leaving users and moderators in the dark." UK regulator Ofcom analysed OnlyFans, Twitch, Snapchat, TikTok, Brand New Tube and BitChute's terms and found each of them was "difficult to read" and would take up to an hour to get through. They also noted that few provide detailed information about content that violates the terms or penalties for breaking the rules.

Helping users better understand platforms' terms of service was one of the key themes at last month's Trust and Safety hackathon attended by dozens of

trust and safety workers ahead of TrustCon (I was there too). If I was one of the video-sharing platforms mentioned in the report, I'd be taking a closer look.

New research looking at Facebook's interventions during the 2020 US election "highlight the fallibility of corporate collaboration and the need for government **transparency** mandates", according to an analysis by Justin Hendrix, CEO and editor of *Tech Policy Press*, and Paul Barrett, deputy director at the Center for Business and Human Rights at NYU Stern. Looking at the papers published in Science and Nature as well as media coverage, the pair raised questions about whether "Meta may be keen to show that its model in this collaboration can inform such rules" as the Digital Services Act and proposed legislation in the US. My read of the week.

Also in this section...

- The Internet Speech Case That the Supreme Court Can't Dodge (*Wired*)
- Coming regulations mean game developers must be proactive on safety and trust (*Venturebeat*)
- The Lawfare Podcast: Can We Build a Trustworthy Future Web? (*Lawfare*)
- France: Proposed internet bill threatens online speech (*Article19*)
- EU tech envoy: 'The winds have changed' on regulating Silicon Valley (*Politico*)

Products

Features, functionality and technology shaping online speech

TikTok users will soon be able to opt out of a **personalised feed** as part of the platform's effort to comply with the Digital Services Act. According to a blog post, users can choose to be served popular content from where they live and to have chronological content from Following and Friends. It will also update its reporting flows to allow users to flag content they believe is illegal, another DSA requirement, as well as "provide our community in Europe with information about a broader range of content moderation decisions"; however, details about how this works in practice are scarce.

Become an EiM member
Everything in Moderation is your guide to understanding how content moderation is changing the world.

Every week, industry experts and leaders read the weekly newsletter from start to end. Here are some nice things they've said:
- "*If you care about any of this and are not subscribing, you should!*" - Ashley
- "*He doesn't miss the little-covered but important stories*" - Kevin
- "*Every edition should be considered must-read content for those in the trust & safety space*" - Justin

Becoming a member helps me continue to curate the information and ideas you need to make the internet a safer, better place for everyone. Without its current members, EiM would have folded by now. So thanks to all members, past, current and future — BW

Platforms

Social networks and the application of content guidelines

Former moderators working for Sama on **OpenAI**'s ChatGPT output have spoken out about their working conditions and the efforts of their job reviewing and testing large language models. This report from *The Guardian* quotes two moderators and explains the fallout of reading violent and sexualised text snippets all day: group calls to swap horror stories, hobbies falling by the wayside and couples splitting up. Back in July, four moderators filed a petition to Kenya's National Assembly calling for an investigation into the working conditions of outsourced workers on AI content moderation.

Also in this section...

- Twitch expands its ban on gambling livestreams (*NBC News*)
- Musk says X to fund legal bills if users 'unfairly treated' over posts (*Al Jazeera*)
- Accenture: tech's silent partner cuts deep in Ireland (*The Times*)
- X tries to win back advertisers with brand safety promises (*The Register*)

People

Those impacting the future of online safety and moderation

The Center for Countering Hate and its research output has attracted significant media coverage since its founding in 2018. It has shone a light on platforms' failures to stop the spread of hate (EiM #124) so it was a matter of time before Elon Musk took aim at the British-based NGO.

Twitter/X's CEO called the organisation "evil" but also went after its founder and CEO **Imran Ahmed**, who he called a "rat". Ahmed is a former senior political advisor for the Labour party and, before that worked for Merrill Lunch and as a strategy consultant for the best part of a decade.

He writes in *The Guardian* this week that "attacking CCDH will not remove the neo-Nazis, white supremacists and disinformation superspreaders he has allowed back". And he warns that "If we allow ourselves to be strongarmed by Musk and X, it will give the green light to every other social media behemoth to do the same to anyone who dares to speak truth to power.".

Posts of note

Handpicked posts that caught my eye this week

- "What about the Trust and Safety Council we've asked her to reinstate?!" - **Global Project Against Hate and Extremism** wonders whether X's reformed 'client council' is a sign of things to come (narrator: it's not)
- "These models will cause big Online Safety Bill headaches." - **Matthew Feeney**, head of technology and innovation at CPS Think Tank, on new research that shows AI models have political biases.
- "AI generated child sex images: a "nightmare" "worst case scenario" for trust and safety? I argue otherwise" - former Protasia Foundation exec director and trust and safety consultant **Jeremy Malcolm** with an interesting thread.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 2000+ EiM subscribers.

Roblox is looking for a Product Marketing Lead, Trust and Safety.

The role is a partner for its Safety Product Management teams —who are responsible for Roblox's policies and scaled operations— and will be responsible for positioning, messaging, content creation, channel planning and measurement for the Safety group's product and policy launches.

It's an interesting role as communications becomes recognised as a strategic trust & safety tool (c/o Mike Masnick over at *Techdirt*). Whoever is successful will have 7+ years of product marketing experience with a leading global technology platform, with experience in trust and safety, privacy, or policy preferred.

The role is based at Roblox HQ in San Mateo, California with a not insignificant (and oddly specific) salary of $212,530—260,790.

Enjoy today's edition Michael? *If so...*

1. ***Hit the thumbs*** *or send a quick reply to share feedback or just say hello*
2. ***Forward to a friend*** *or colleague (PS they can sign up here)*
3. ***Become an EiM member*** *and get access to original analysis and the whole EiM archive for less than the price of a coffee a week*

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: noreply=everythinginmoderation.co@m.ghost.io on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, 11 August 2023 10:01 PM
To: s 47E(c), s 47E
Subject: 'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

By Ben Whitelaw • 11 Aug 2023 •

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

This edition comes after a two-week break due to a holiday and then an unexpected hospital trip last Friday. Everyone and everything is fine but I'm sorry to those of you who were expecting their weekly newsletter to arrive and who were surprised when it didn't.

Today's newsletter shows, once again, the very tangible impact of incoming online speech regulation, both in the ways that it is shaping platform features as well as how it will shape future legislation.

Before we get into the big stories, I want to welcome a host of new subscribers since the last EiM from Google, Checkstep, Uber, TrustLab, Electronic Arts, Feeld, Mailchimp, Bumble, Shopify, Etsy and others. Your feedback is key to the newsletter hitting the mark each week. Email me or hit the thumbs at the end of today's edition.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

A new report has found that the **terms and conditions** of video-sharing platforms are "lengthy, impenetrable and, in some cases, inconsistent" and "risk leaving users and moderators in the dark." UK regulator Ofcom analysed OnlyFans, Twitch, Snapchat, TikTok, Brand New Tube and BitChute's terms and found each of them was "difficult to read" and would take up to an hour to get through. They also noted that few provide detailed information about content that violates the terms or penalties for breaking the rules.

Helping users better understand platforms' terms of service was one of the key themes at last month's Trust and Safety hackathon attended by dozens of

trust and safety workers ahead of TrustCon (I was there too). If I was one of the video-sharing platforms mentioned in the report, I'd be taking a closer look.

New research looking at Facebook's interventions during the 2020 US election "highlight the fallibility of corporate collaboration and the need for government **transparency** mandates", according to an analysis by Justin Hendrix, CEO and editor of *Tech Policy Press*, and Paul Barrett, deputy director at the Center for Business and Human Rights at NYU Stern. Looking at the papers published in Science and Nature as well as media coverage, the pair raised questions about whether "Meta may be keen to show that its model in this collaboration can inform such rules" as the Digital Services Act and proposed legislation in the US. My read of the week.

Also in this section...

- The Internet Speech Case That the Supreme Court Can't Dodge (*Wired*)
- Coming regulations mean game developers must be proactive on safety and trust (*Venturebeat*)
- The Lawfare Podcast: Can We Build a Trustworthy Future Web? (*Lawfare*)
- France: Proposed internet bill threatens online speech (*Article19*)
- EU tech envoy: 'The winds have changed' on regulating Silicon Valley (*Politico*)

Products

Features, functionality and technology shaping online speech

TikTok users will soon be able to opt out of a **personalised feed** as part of the platform's effort to comply with the Digital Services Act. According to a blog post, users can choose to be served popular content from where they live and to have chronological content from Following and Friends. It will also update its reporting flows to allow users to flag content they believe is illegal, another DSA requirement, as well as "provide our community in Europe with information about a broader range of content moderation decisions"; however, details about how this works in practice are scarce.

Become an EiM member
Everything in Moderation is your guide to understanding how content moderation is changing the world.

Every week, industry experts and leaders read the weekly newsletter from start to end. Here are some nice things they've said:
- "*If you care about any of this and are not subscribing, you should!*" - Ashley
- "*He doesn't miss the little-covered but important stories*" - Kevin
- "*Every edition should be considered must-read content for those in the trust & safety space*" - Justin

Becoming a member helps me continue to curate the information and ideas you need to make the internet a safer, better place for everyone. Without its current members, EiM would have folded by now. So thanks to all members, past, current and future — BW

Platforms

Social networks and the application of content guidelines

Former moderators working for Sama on **OpenAI**'s ChatGPT output have spoken out about their working conditions and the efforts of their job reviewing and testing large language models. This report from *The Guardian* quotes two moderators and explains the fallout of reading violent and sexualised text snippets all day: group calls to swap horror stories, hobbies falling by the wayside and couples splitting up. Back in July, four moderators filed a petition to Kenya's National Assembly calling for an investigation into the working conditions of outsourced workers on AI content moderation.

Also in this section...

- Twitch expands its ban on gambling livestreams (*NBC News*)
- Musk says X to fund legal bills if users 'unfairly treated' over posts (*Al Jazeera*)
- Accenture: tech's silent partner cuts deep in Ireland (*The Times*)
- X tries to win back advertisers with brand safety promises (*The Register*)

People

Those impacting the future of online safety and moderation

The Center for Countering Hate and its research output has attracted significant media coverage since its founding in 2018. It has shone a light on platforms' failures to stop the spread of hate (EiM #124) so it was a matter of time before Elon Musk took aim at the British-based NGO.

Twitter/X's CEO called the organisation "evil" but also went after its founder and CEO **Imran Ahmed**, who he called a "rat". Ahmed is a former senior political advisor for the Labour party and, before that worked for Merrill Lunch and as a strategy consultant for the best part of a decade.

He writes in *The Guardian* this week that "attacking CCDH will not remove the neo-Nazis, white supremacists and disinformation superspreaders he has allowed back". And he warns that "If we allow ourselves to be strongarmed by Musk and X, it will give the green light to every other social media behemoth to do the same to anyone who dares to speak truth to power.".

Posts of note

Handpicked posts that caught my eye this week

- "What about the Trust and Safety Council we've asked her to reinstate?!" - **Global Project Against Hate and Extremism** wonders whether X's reformed 'client council' is a sign of things to come (narrator: it's not)
- "These models will cause big Online Safety Bill headaches." - **Matthew Feeney**, head of technology and innovation at CPS Think Tank, on new research that shows AI models have political biases.
- "AI generated child sex images: a "nightmare" "worst case scenario" for trust and safety? I argue otherwise" - former Protasia Foundation exec director and trust and safety consultant **Jeremy Malcolm** with an interesting thread.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 2000+ EiM subscribers.

Roblox is looking for a Product Marketing Lead, Trust and Safety.

The role is a partner for its Safety Product Management teams —who are responsible for Roblox's policies and scaled operations— and will be responsible for positioning, messaging, content creation, channel planning and measurement for the Safety group's product and policy launches.

It's an interesting role as communications becomes recognised as a strategic trust & safety tool (c/o Mike Masnick over at *Techdirt*). Whoever is successful will have 7+ years of product marketing experience with a leading global technology platform, with experience in trust and safety, privacy, or policy preferred.

The role is based at Roblox HQ in San Mateo, California with a not insignificant (and oddly specific) salary of $212,530—260,790.

Enjoy today's edition ? If so...

1. ***Hit the thumbs*** *or send a quick reply to share feedback or just say hello*
2. ***Forward to a friend*** *or colleague (PS they can sign up here)*
3. ***Become an EiM member*** *and get access to original analysis and the whole EiM archive for less than the price of a coffee a week*

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: noreply=everythinginmoderation.co@m.ghost.io on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, 11 August 2023 10:01 PM
To: s 47E(c), s 47E
Subject: 'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

By Ben Whitelaw • 11 Aug 2023 •

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

This edition comes after a two-week break due to a holiday and then an unexpected hospital trip last Friday. Everyone and everything is fine but I'm sorry to those of you who were expecting their weekly newsletter to arrive and who were surprised when it didn't.

Today's newsletter shows, once again, the very tangible impact of incoming online speech regulation, both in the ways that it is shaping platform features as well as how it will shape future legislation.

Before we get into the big stories, I want to welcome a host of new subscribers since the last EiM from Google, Checkstep, Uber, TrustLab, Electronic Arts, Feeld, Mailchimp, Bumble, Shopify, Etsy and others. Your feedback is key to the newsletter hitting the mark each week. Email me or hit the thumbs at the end of today's edition.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

A new report has found that the **terms and conditions** of video-sharing platforms are "lengthy, impenetrable and, in some cases, inconsistent" and "risk leaving users and moderators in the dark." UK regulator Ofcom analysed OnlyFans, Twitch, Snapchat, TikTok, Brand New Tube and BitChute's terms and found each of them was "difficult to read" and would take up to an hour to get through. They also noted that few provide detailed information about content that violates the terms or penalties for breaking the rules.

Helping users better understand platforms' terms of service was one of the key themes at last month's Trust and Safety hackathon attended by dozens of

trust and safety workers ahead of TrustCon (I was there too). If I was one of the video-sharing platforms mentioned in the report, I'd be taking a closer look.

New research looking at Facebook's interventions during the 2020 US election "highlight the fallibility of corporate collaboration and the need for government **transparency** mandates", according to an analysis by Justin Hendrix, CEO and editor of *Tech Policy Press*, and Paul Barrett, deputy director at the Center for Business and Human Rights at NYU Stern. Looking at the papers published in Science and Nature as well as media coverage, the pair raised questions about whether "Meta may be keen to show that its model in this collaboration can inform such rules" as the Digital Services Act and proposed legislation in the US. My read of the week.

Also in this section...

- The Internet Speech Case That the Supreme Court Can't Dodge (*Wired*)
- Coming regulations mean game developers must be proactive on safety and trust (*Venturebeat*)
- The Lawfare Podcast: Can We Build a Trustworthy Future Web? (*Lawfare*)
- France: Proposed internet bill threatens online speech (*Article19*)
- EU tech envoy: 'The winds have changed' on regulating Silicon Valley (*Politico*)

Products

Features, functionality and technology shaping online speech

TikTok users will soon be able to opt out of a **personalised feed** as part of the platform's effort to comply with the Digital Services Act. According to a blog post, users can choose to be served popular content from where they live and to have chronological content from Following and Friends. It will also update its reporting flows to allow users to flag content they believe is illegal, another DSA requirement, as well as "provide our community in Europe with information about a broader range of content moderation decisions"; however, details about how this works in practice are scarce.

Become an EiM member
Everything in Moderation is your guide to understanding how content moderation is changing the world.

Every week, industry experts and leaders read the weekly newsletter from start to end. Here are some nice things they've said:
- "*If you care about any of this and are not subscribing, you should!*" - Ashley
- "*He doesn't miss the little-covered but important stories*" - Kevin
- "*Every edition should be considered must-read content for those in the trust & safety space*" - Justin

Becoming a member helps me continue to curate the information and ideas you need to make the internet a safer, better place for everyone. Without its current members, EiM would have folded by now. So thanks to all members, past, current and future — BW

Platforms

Social networks and the application of content guidelines

Former moderators working for Sama on **OpenAI**'s ChatGPT output have spoken out about their working conditions and the efforts of their job reviewing and testing large language models. This report from *The Guardian* quotes two moderators and explains the fallout of reading violent and sexualised text snippets all day: group calls to swap horror stories, hobbies falling by the wayside and couples splitting up. Back in July, four moderators filed a petition to Kenya's National Assembly calling for an investigation into the working conditions of outsourced workers on AI content moderation.

Also in this section...

- Twitch expands its ban on gambling livestreams (*NBC News*)
- Musk says X to fund legal bills if users 'unfairly treated' over posts (*Al Jazeera*)
- Accenture: tech's silent partner cuts deep in Ireland (*The Times*)
- X tries to win back advertisers with brand safety promises (*The Register*)

People

Those impacting the future of online safety and moderation

The Center for Countering Hate and its research output has attracted significant media coverage since its founding in 2018. It has shone a light on platforms' failures to stop the spread of hate (EiM #124) so it was a matter of time before Elon Musk took aim at the British-based NGO.

Twitter/X's CEO called the organisation "evil" but also went after its founder and CEO **Imran Ahmed**, who he called a "rat". Ahmed is a former senior political advisor for the Labour party and, before that worked for Merrill Lunch and as a strategy consultant for the best part of a decade.

He writes in *The Guardian* this week that "attacking CCDH will not remove the neo-Nazis, white supremacists and disinformation superspreaders he has allowed back". And he warns that "If we allow ourselves to be strongarmed by Musk and X, it will give the green light to every other social media behemoth to do the same to anyone who dares to speak truth to power.".

Posts of note

Handpicked posts that caught my eye this week

- "What about the Trust and Safety Council we've asked her to reinstate?!" - **Global Project Against Hate and Extremism** wonders whether X's reformed 'client council' is a sign of things to come (narrator: it's not)
- "These models will cause big Online Safety Bill headaches." - **Matthew Feeney**, head of technology and innovation at CPS Think Tank, on new research that shows AI models have political biases.
- "AI generated child sex images: a "nightmare" "worst case scenario" for trust and safety? I argue otherwise" - former Protasia Foundation exec director and trust and safety consultant **Jeremy Malcolm** with an interesting thread.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 2000+ EiM subscribers.

Roblox is looking for a Product Marketing Lead, Trust and Safety.

The role is a partner for its Safety Product Management teams —who are responsible for Roblox's policies and scaled operations— and will be responsible for positioning, messaging, content creation, channel planning and measurement for the Safety group's product and policy launches.

It's an interesting role as communications becomes recognised as a strategic trust & safety tool (c/o Mike Masnick over at *Techdirt*). Whoever is successful will have 7+ years of product marketing experience with a leading global technology platform, with experience in trust and safety, privacy, or policy preferred.

The role is based at Roblox HQ in San Mateo, California with a not insignificant (and oddly specific) salary of $212,530—260,790.

Enjoy today's edition Elizabeth? *If so...*

1. ***Hit the thumbs*** *or send a quick reply to share feedback or just say hello*
2. ***Forward to a friend*** *or colleague (PS they can sign up here)*
3. ***Become an EiM member*** *and get access to original analysis and the whole EiM archive for less than the price of a coffee a week*

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: noreply=everythinginmoderation.co@m.ghost.io on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, 11 August 2023 10:01 PM
To: s 47E(c), s 47F
Subject: 'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

By Ben Whitelaw • 11 Aug 2023 •

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

This edition comes after a two-week break due to a holiday and then an unexpected hospital trip last Friday. Everyone and everything is fine but I'm sorry to those of you who were expecting their weekly newsletter to arrive and who were surprised when it didn't.

Today's newsletter shows, once again, the very tangible impact of incoming online speech regulation, both in the ways that it is shaping platform features as well as how it will shape future legislation.

Before we get into the big stories, I want to welcome a host of new subscribers since the last EiM from Google, Checkstep, Uber, TrustLab, Electronic Arts, Feeld, Mailchimp, Bumble, Shopify, Etsy and others. Your feedback is key to the newsletter hitting the mark each week. Email me or hit the thumbs at the end of today's edition.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

A new report has found that the **terms and conditions** of video-sharing platforms are "lengthy, impenetrable and, in some cases, inconsistent" and "risk leaving users and moderators in the dark." UK regulator Ofcom analysed OnlyFans, Twitch, Snapchat, TikTok, Brand New Tube and BitChute's terms and found each of them was "difficult to read" and would take up to an hour to get through. They also noted that few provide detailed information about content that violates the terms or penalties for breaking the rules.

Helping users better understand platforms' terms of service was one of the key themes at last month's Trust and Safety hackathon attended by dozens of

trust and safety workers ahead of TrustCon (I was there too). If I was one of the video-sharing platforms mentioned in the report, I'd be taking a closer look.

New research looking at Facebook's interventions during the 2020 US election "highlight the fallibility of corporate collaboration and the need for government **transparency** mandates", according to an analysis by Justin Hendrix, CEO and editor of *Tech Policy Press*, and Paul Barrett, deputy director at the Center for Business and Human Rights at NYU Stern. Looking at the papers published in Science and Nature as well as media coverage, the pair raised questions about whether "Meta may be keen to show that its model in this collaboration can inform such rules" as the Digital Services Act and proposed legislation in the US. My read of the week.

Also in this section...

- The Internet Speech Case That the Supreme Court Can't Dodge (*Wired*)
- Coming regulations mean game developers must be proactive on safety and trust (*Venturebeat*)
- The Lawfare Podcast: Can We Build a Trustworthy Future Web? (*Lawfare*)
- France: Proposed internet bill threatens online speech (*Article19*)
- EU tech envoy: 'The winds have changed' on regulating Silicon Valley (*Politico*)

Products

Features, functionality and technology shaping online speech

TikTok users will soon be able to opt out of a **personalised feed** as part of the platform's effort to comply with the Digital Services Act. According to a blog post, users can choose to be served popular content from where they live and to have chronological content from Following and Friends. It will also update its reporting flows to allow users to flag content they believe is illegal, another DSA requirement, as well as "provide our community in Europe with information about a broader range of content moderation decisions"; however, details about how this works in practice are scarce.

Become an EiM member
Everything in Moderation is your guide to understanding how content moderation is changing the world.

Every week, industry experts and leaders read the weekly newsletter from start to end. Here are some nice things they've said:
- "*If you care about any of this and are not subscribing, you should!*" - Ashley
- "*He doesn't miss the little-covered but important stories*" - Kevin
- "*Every edition should be considered must-read content for those in the trust & safety space*" - Justin

Becoming a member helps me continue to curate the information and ideas you need to make the internet a safer, better place for everyone. Without its current members, EiM would have folded by now. So thanks to all members, past, current and future — BW

Platforms

Social networks and the application of content guidelines

Former moderators working for Sama on **OpenAI**'s ChatGPT output have spoken out about their working conditions and the efforts of their job reviewing and testing large language models. This report from *The Guardian* quotes two moderators and explains the fallout of reading violent and sexualised text snippets all day: group calls to swap horror stories, hobbies falling by the wayside and couples splitting up. Back in July, four moderators filed a petition to Kenya's National Assembly calling for an investigation into the working conditions of outsourced workers on AI content moderation.

Also in this section...

- Twitch expands its ban on gambling livestreams (*NBC News*)
- Musk says X to fund legal bills if users 'unfairly treated' over posts (*Al Jazeera*)
- Accenture: tech's silent partner cuts deep in Ireland (*The Times*)
- X tries to win back advertisers with brand safety promises (*The Register*)

People

Those impacting the future of online safety and moderation

The Center for Countering Hate and its research output has attracted significant media coverage since its founding in 2018. It has shone a light on platforms' failures to stop the spread of hate (EiM #124) so it was a matter of time before Elon Musk took aim at the British-based NGO.

Twitter/X's CEO called the organisation "evil" but also went after its founder and CEO **Imran Ahmed**, who he called a "rat". Ahmed is a former senior political advisor for the Labour party and, before that worked for Merrill Lunch and as a strategy consultant for the best part of a decade.

He writes in *The Guardian* this week that "attacking CCDH will not remove the neo-Nazis, white supremacists and disinformation superspreaders he has allowed back". And he warns that "If we allow ourselves to be strongarmed by Musk and X, it will give the green light to every other social media behemoth to do the same to anyone who dares to speak truth to power.".

Posts of note

Handpicked posts that caught my eye this week

- "What about the Trust and Safety Council we've asked her to reinstate?!" - **Global Project Against Hate and Extremism** wonders whether X's reformed 'client council' is a sign of things to come (narrator: it's not)
- "These models will cause big Online Safety Bill headaches." - **Matthew Feeney**, head of technology and innovation at CPS Think Tank, on new research that shows AI models have political biases.
- "AI generated child sex images: a "nightmare" "worst case scenario" for trust and safety? I argue otherwise" - former Protasia Foundation exec director and trust and safety consultant **Jeremy Malcolm** with an interesting thread.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 2000+ EiM subscribers.

Roblox is looking for a Product Marketing Lead, Trust and Safety.

The role is a partner for its Safety Product Management teams —who are responsible for Roblox's policies and scaled operations— and will be responsible for positioning, messaging, content creation, channel planning and measurement for the Safety group's product and policy launches.

It's an interesting role as communications becomes recognised as a strategic trust & safety tool (c/o Mike Masnick over at *Techdirt*). Whoever is successful will have 7+ years of product marketing experience with a leading global technology platform, with experience in trust and safety, privacy, or policy preferred.

The role is based at Roblox HQ in San Mateo, California with a not insignificant (and oddly specific) salary of $212,530—260,790.

Enjoy today's edition Phoebe? *If so...*

1. ***Hit the thumbs*** *or send a quick reply to share feedback or just say hello*
2. ***Forward to a friend*** *or colleague (PS they can sign up here)*
3. ***Become an EiM member*** *and get access to original analysis and the whole EiM archive for less than the price of a coffee a week*

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

From: noreply=everythinginmoderation.co@m.ghost.io on behalf of Ben from Everything in Moderation* <noreply@everythinginmoderation.co>
Sent: Friday, 11 August 2023 10:01 PM
To: s 47E(c), s 47E
Subject: 'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

EVERYTHING IN MODERATION*

'Impenetrable' platform rules, TikTok shares DSA compliance and the other CEO fighting Musk

By Ben Whitelaw • 11 Aug 2023 •

View in browser

Hello and welcome to **Everything in Moderation**, your guide to the policies, products, platforms and people shaping the future of online speech and the internet. It's written by me, Ben Whitelaw and supported by members like you.

This edition comes after a two-week break due to a holiday and then an unexpected hospital trip last Friday. Everyone and everything is fine but I'm sorry to those of you who were expecting their weekly newsletter to arrive and who were surprised when it didn't.

Today's newsletter shows, once again, the very tangible impact of incoming online speech regulation, both in the ways that it is shaping platform features as well as how it will shape future legislation.

Before we get into the big stories, I want to welcome a host of new subscribers since the last EiM from Google, Checkstep, Uber, TrustLab, Electronic Arts, Feeld, Mailchimp, Bumble, Shopify, Etsy and others. Your feedback is key to the newsletter hitting the mark each week. Email me or hit the thumbs at the end of today's edition.

Here's everything in moderation from the last seven days — BW

Policies

New and emerging internet policy and online speech regulation

A new report has found that the **terms and conditions** of video-sharing platforms are "lengthy, impenetrable and, in some cases, inconsistent" and "risk leaving users and moderators in the dark." UK regulator Ofcom analysed OnlyFans, Twitch, Snapchat, TikTok, Brand New Tube and BitChute's terms and found each of them was "difficult to read" and would take up to an hour to get through. They also noted that few provide detailed information about content that violates the terms or penalties for breaking the rules.

Helping users better understand platforms' terms of service was one of the key themes at last month's Trust and Safety hackathon attended by dozens of

trust and safety workers ahead of TrustCon (I was there too). If I was one of the video-sharing platforms mentioned in the report, I'd be taking a closer look.

New research looking at Facebook's interventions during the 2020 US election "highlight the fallibility of corporate collaboration and the need for government **transparency** mandates", according to an analysis by Justin Hendrix, CEO and editor of *Tech Policy Press*, and Paul Barrett, deputy director at the Center for Business and Human Rights at NYU Stern. Looking at the papers published in Science and Nature as well as media coverage, the pair raised questions about whether "Meta may be keen to show that its model in this collaboration can inform such rules" as the Digital Services Act and proposed legislation in the US. My read of the week.

Also in this section...

- The Internet Speech Case That the Supreme Court Can't Dodge (*Wired*)
- Coming regulations mean game developers must be proactive on safety and trust (*Venturebeat*)
- The Lawfare Podcast: Can We Build a Trustworthy Future Web? (*Lawfare*)
- France: Proposed internet bill threatens online speech (*Article19*)
- EU tech envoy: 'The winds have changed' on regulating Silicon Valley (*Politico*)

Products

Features, functionality and technology shaping online speech

TikTok users will soon be able to opt out of a **personalised feed** as part of the platform's effort to comply with the Digital Services Act. According to a blog post, users can choose to be served popular content from where they live and to have chronological content from Following and Friends. It will also update its reporting flows to allow users to flag content they believe is illegal, another DSA requirement, as well as "provide our community in Europe with information about a broader range of content moderation decisions"; however, details about how this works in practice are scarce.

Become an EiM member
Everything in Moderation is your guide to understanding how content moderation is changing the world.

Every week, industry experts and leaders read the weekly newsletter from start to end. Here are some nice things they've said:
- "*If you care about any of this and are not subscribing, you should!*" - Ashley
- "*He doesn't miss the little-covered but important stories*" - Kevin
- "*Every edition should be considered must-read content for those in the trust & safety space*" - Justin

Becoming a member helps me continue to curate the information and ideas you need to make the internet a safer, better place for everyone. Without its current members, EiM would have folded by now. So thanks to all members, past, current and future — BW

Platforms

Social networks and the application of content guidelines

Former moderators working for Sama on **OpenAI**'s ChatGPT output have spoken out about their working conditions and the efforts of their job reviewing and testing large language models. This report from *The Guardian* quotes two moderators and explains the fallout of reading violent and sexualised text snippets all day: group calls to swap horror stories, hobbies falling by the wayside and couples splitting up. Back in July, four moderators filed a petition to Kenya's National Assembly calling for an investigation into the working conditions of outsourced workers on AI content moderation.

Also in this section...

- Twitch expands its ban on gambling livestreams (*NBC News*)
- Musk says X to fund legal bills if users 'unfairly treated' over posts (*Al Jazeera*)
- Accenture: tech's silent partner cuts deep in Ireland (*The Times*)
- X tries to win back advertisers with brand safety promises (*The Register*)

People

Those impacting the future of online safety and moderation

The Center for Countering Hate and its research output has attracted significant media coverage since its founding in 2018. It has shone a light on platforms' failures to stop the spread of hate (EiM #124) so it was a matter of time before Elon Musk took aim at the British-based NGO.

Twitter/X's CEO called the organisation "evil" but also went after its founder and CEO **Imran Ahmed**, who he called a "rat". Ahmed is a former senior political advisor for the Labour party and, before that worked for Merrill Lunch and as a strategy consultant for the best part of a decade.

He writes in *The Guardian* this week that "attacking CCDH will not remove the neo-Nazis, white supremacists and disinformation superspreaders he has allowed back". And he warns that "If we allow ourselves to be strongarmed by Musk and X, it will give the green light to every other social media behemoth to do the same to anyone who dares to speak truth to power.".

Posts of note

Handpicked posts that caught my eye this week

- "What about the Trust and Safety Council we've asked her to reinstate?!" - **Global Project Against Hate and Extremism** wonders whether X's reformed 'client council' is a sign of things to come (narrator: it's not)
- "These models will cause big Online Safety Bill headaches." - **Matthew Feeney**, head of technology and innovation at CPS Think Tank, on new research that shows AI models have political biases.
- "AI generated child sex images: a "nightmare" "worst case scenario" for trust and safety? I argue otherwise" - former Protasia Foundation exec director and trust and safety consultant **Jeremy Malcolm** with an interesting thread.

Job of the week

Share and discover jobs in trust and safety, content moderation and online safety. Become an EiM member to share your job ad for free with 2000+ EiM subscribers.

Roblox is looking for a Product Marketing Lead, Trust and Safety.

The role is a partner for its Safety Product Management teams —who are responsible for Roblox's policies and scaled operations— and will be responsible for positioning, messaging, content creation, channel planning and measurement for the Safety group's product and policy launches.

It's an interesting role as communications becomes recognised as a strategic trust & safety tool (c/o Mike Masnick over at *Techdirt*). Whoever is successful will have 7+ years of product marketing experience with a leading global technology platform, with experience in trust and safety, privacy, or policy preferred.

The role is based at Roblox HQ in San Mateo, California with a not insignificant (and oddly specific) salary of $212,530—260,790.

Enjoy today's edition Ahram? *If so...*

1. ***Hit the thumbs*** *or send a quick reply to share feedback or just say hello*
2. ***Forward to a friend*** *or colleague (PS they can sign up here)*
3. ***Become an EiM member*** *and get access to original analysis and the whole EiM archive for less than the price of a coffee a week*

Everything in Moderation *is a weekly newsletter about content moderation and the policies, products, platforms and people shaping its future.*

Name	Position title
Julie Inman-Grant	eSafety Commissioner
s 47E(c), s 47F	