ISD | Powering solutions to extremism and polarisation

Gwead Gwamal y We

Y dirwedd gydgysylltiedig ar-lein o iaith casineb, eithafiaeth, terfysgaeth a mudiadau cynllwyn niweidiol yn y Deyrnas Unedig

Milo Comerford, Jacob Davey, Jakob Guhl

Ynghylch yr adroddiad hwn

Mae'r adroddiad hwn yn rhoi cipolwg ar y dirwedd ar-lein o ran terfysgaeth, eithafiaeth ac iaith casineb sy'n gysylltiedig â'r Deyrnas Unedig. Daw'r adroddiad yn sgil y dadansoddiad digidol a gynhaliwyd gan Institute for Strategic Dialogue (ISD) a CASM Technology ar gyfer Ofcom. Mae'n tynnu sylw at dueddiadau sy'n ymwneud â gwahanol gyfranogwyr sy'n hyrwyddo eithafiaeth, iaith casineb a damcaniaethau cynllwyn niweidiol ar draws nifer o lwyfannau cyfryngau cymdeithasol perthnasol. Mae'r llwyfannau hyn yn cynnwys Facebook, YouTube, Twitter, Instagram, Reddit, Telegram a 4chan. Defnyddiodd ymchwilwyr ISD fethodolegau ansoddol a meintiol i greu'r ymchwil hon, sy'n seiliedig ar system dadansoddi data unigryw o'r enw "Beam". Mae Beam yn defnyddio dysgu peirianyddol a phrosesu iaith naturiol er mwyn cynnal gwaith ymchwil arloesol ar y cyfryngau cymdeithasol ar draws llwyfannau.

ISD | Powering solutions to extremism and polarisation

Amman | Berlin | Llundain | Paris | Washington DC

CYNNWYS

ISD | Powering solutions to extremism, hate and disinformation

Crynodeb Gweithredol

Mae pwysigrwydd cynyddol llwyfannau cyfryngau cymdeithasol wedi siapio cymdeithas mewn ffyrdd dwys. Mae'r llwyfannau hyn wedi rhoi'r gallu i ddefnyddwyr fynegi eu hunain yn rhydd, adeiladu cymunedau, ac ymgysylltu ag ystod eang o safbwyntiau. Ar yr un pryd, mae tirwedd niwed ar-lein yn newid o hyd, gyda chyfryngau cymdeithasol yn cael eu camddefnyddio mewn modd mwy soffistigedig byth er mwyn annog casineb, lledaenu damcaniaethau cynllwyn ac ysgogi trais go iawn yn y byd. Er bod gwaith ymchwil gan Ofcom yn dangos nad yw'r rhan fwyaf o ddefnyddwyr yn dod ar draws cynnwys sy'n atgas, yn dreisgar, neu'n cynnwys camwybodaeth yn aml, mae angen cymryd y bygythiadau hyn o ddifrif o ystyried eu potensial i achosi niwed go iawn yn y byd.[i]

Yn y Deyrnas Unedig, mae cyfryngau cymdeithasol wedi chwarae rhan gynyddol mewn symudiadau eithafwyr treisgar ar draws y sbectrwm ideolegol.[ii] Yn y cyfamser, yn sgil y pandemig Covid-19, mae llwyfannau cyfryngau cymdeithasol wedi gweld cynnydd mewn achosion o gasineb sy'n targedu cymunedau agored i niwed. Mae hyn hefyd yn wir am achosion o gam-drin ac aflonyddu yn erbyn ffigurau cyhoeddus, fel gweithwyr iechyd, newyddiadurwyr a swyddogion etholedig.[iii]

Mae niwed o'r fath yn cael ei sbarduno gan gontinwwm o weithredwyr eithafol sydd wedi'u trefnu a rhwydweithiau llacach sy'n gysylltiedig â chasineb gwrth-leiafrifol a hyrwyddo damcaniaethau cynllwyn.[1] Mae unigolion yn y mudiadau hyn yn cael eu diffinio fwy gan yr is-ddiwylliannau ar-lein y maent yn byw ynddynt, ac yn aml yn symud yn rhwydd rhyngddynt, na'u haelodaeth o sefydliadau terfysgol gwaharddedig. Mae camwybodaeth, damcaniaethau cynllwyn, iaith casineb, aflonyddu ac eithafiaeth dreisgar yn aml yn cydblethu mewn ffyrdd sy'n anodd iawn eu gwahanu a'u hynysu.

Bydd deall pa mor aneglur yw'r llinellau hyn yn ffactor pwysig wrth ddylunio ymatebion polisi a rheoleiddiol effeithiol i weithgarwch anghyfreithlon a niweidiol gan fod y rhain yn dibynnu ar ddiffiniadau cyfreithiol clir. Mae'r cyfnod hwn o newid yn y dirwedd ar-lein yn cyd-daro â chyfnod lle mae llunwyr polisïau yn y Deyrnas Unedig, Ewrop a ledled y byd yn mynd ati i fynd i'r afael â'r niwed seicolegol a chorfforol sy'n gysylltiedig â gweithgarwch cyfryngau cymdeithasol. Wrth i fframweithiau rheoleiddio gael eu datblygu, mae angen dybryd am well tystiolaeth ynghylch natur y gweithgarwch anghyfreithiol a niweidiol sydd i'w weld ar wahanol lwyfannau cyfryngau cymdeithasol, a'r grwpiau a'r cyfranogwyr sy'n gyfrifol amdano.

Comisiynwyd yr ymchwil hon gan Ofcom i ddarparu tystiolaeth ynghylch ymddangosiad niwed ar-lein ar draws llwyfannau'r dirwedd eang hon, sef materion rhyngberthynol terfysgol, eithafol ac iaith casineb. Ni ddylid ei chymryd fel adlewyrchiad o ddull rheoleiddio Ofcom, ond mae'n rhoi darlun bras o sefyllfa'r bygythiadau hyn yn y Deyrnas Unedig ar hyn o bryd. Mae'n werth tynnu sylw o'r

[1] Mae ISD, fel y diffinnir yn llawn ar dudalen 6, yn diffinio eithafiaeth fel defnyddio trais, gwleidyddiaeth neu newid cymdeithasol i hyrwyddo ideoleg goruchafiaethol, sy'n fframio goroesiad 'grŵp mewnol' drwy ddinistrio 'grŵp allanol'. Yn yr adroddiad hwn rydym yn canolbwyntio'n benodol ar gynnwys eithafol sy'n gysylltiedig ag ysgogi, bygythiadau treisgar neu aflonyddu, sy'n cyfeirio casineb yn erbyn grŵp gwarchodedig, neu sy'n parhau i ledaenu twyllwybodaeth niweidiol. Mae damcaniaethau cynllwyn yn esbonio digwyddiadau o ran grŵp bach o bobl bwerus yn gweithredu'n gyfrinachol er eu budd eu hunain yn erbyn y lles cyffredin. Mae'r adroddiad hwn yn canolbwyntio ar symudiadau cynllwyn sy'n gysylltiedig â niwed yn y byd go iawn, gan gynnwys annog aflonyddu a bygythiadau treisgar, neu gasineb yn erbyn grŵp gwarchodedig.

.

cychwyn cyntaf at y ffaith bod ymchwilwyr a rheoleiddwyr fel ei gilydd yn wynebu heriau mawr wrth geisio deall y materion hyn ar raddfa fawr. Mae llwyfannau'n gosod cyfyngiadau ar fynediad at ddata, ac mae pob llwyfan yn gwneud hynny mewn ffordd wahanol er mwyn ei gwneud yn anodd cymharu pa mor gyffredin yw cynnwys niweidiol ar draws gwahanol lwyfannau. Oherwydd y cyfyngiadau hyn, mae'n anodd iawn i rheoleiddwyr neu ymchwilwyr annibynnol ddarparu asesiad diffiniol o raddfa y math yma o niwed ar-lein.

Dull Gweithredu

Gan ganolbwyntio ar saith llwyfan cyfryngau cymdeithasol yn benodol, sef Facebook, YouTube, Twitter, Reddit, Instagram, Telegram a 4chan, mae'r adroddiad hwn yn dadansoddi gwahanol gymunedau cydgysylltiedig ar-lein sy'n cymryd rhan mewn gweithgarwch anghyfreithlon a niweidiol posib ac yn targedu cynulleidfaoedd yn y Deyrnas Unedig – yn benodol y rhai sy'n hyrwyddo terfysgaeth, eithafiaeth, iaith casineb a damcaniaethau cynllwyn niweidiol. Mae'r ymchwil hon yn dangos tueddiadau eang ar draws is-set o gyfrifon sy'n gysylltiedig â'r ffenomenau hyn – nid yw'n honni ei fod yn mapio holl gyfrifon a sianeli'r DU sy'n hyrwyddo casineb ac eithafiaeth ar draws y llwyfannau hyn.

Mae'r llwyfannau hyn wedi cael eu cynnwys oherwydd bod data eithaf tebyg ar gael (er ei fod o raddau a mathau gwahanol) drwy ryngwynebau rhaglennu cymwysiadau cyhoeddus (APIs), ymchwil flaenorol sy'n nodi eu bod yn lleoliadau perthnasol lle mae eithafwyr wedi ceisio ysgogi,[iv] yn ogystal â'u hamlygrwydd yn y DU. Mae'r holl wasanaethau hyn ymysg y 10 prif lwyfan ar gyfer defnyddwyr yn y DU,[v] ac eithrio'r bwrdd delweddau 4chan, a gafodd ei gynnwys oherwydd rôl bwysig bwrdd /pol/ y llwyfan mewn is-ddiwylliant eithafiaeth adain dde eithafol atgas.[vi]

Nododd arbenigwyr pwnc yn ISD (drwy waith unigolion ac yn lled-awtomatig) 768 o gyfrifon, grwpiau a sianeli a oedd yn bodloni ein diffiniadau gweithredol o derfysgaeth, eithafiaeth, iaith casineb neu ddamcaniaethau cynllwyn niweidiol – a amlinellir yn fanwl isod – ac a oedd yn cael eu hystyried yn berthnasol i'r DU. Roedd y rhain yn cynnwys cyfrifon sy'n gysylltiedig â grŵp neu weithredwr niweidiol hysbys yn y DU, cyfrif cyfryngau cymdeithasol sy'n ymwneud â gweithgarwch niweidiol sy'n canolbwyntio'n benodol ar y DU, neu gymuned niweidiol ar-lein lle mae tystiolaeth sylweddol o ymgysylltiad gan unigolion yn y DU.

Gan amrywio o oruchafiaethwyr gwyn eithafiaeth adain dde i eithafwyr Islamaidd, damcaniaethwyr cynllwyn gwrth-semitaidd i grwpiau gwrth-Foslemaidd Hindutva, cafodd y cyfrifon a'r sianeli hyn eu codio yn ôl eu cefnogaeth i derfysgaeth, gweithgarwch eithafol, targedu grŵp gwarchodedig mewn ffordd atgas, neu ledaenu cynllwynion sy'n gysylltiedig â niwed go iawn yn y byd.

Wrth gasglu data o'r cyfrifon hyn rhwng 1 Hydref 2021 a 31 Mawrth 2022, fe wnaethom gasglu ychydig dros 2.5 miliwn o negeseuon. Er nad oedd yr holl gynnwys a rannwyd gan y cyfrifon hyn yn amlwg yn anghyfreithlon nac yn niweidiol, defnyddiodd ymchwilwyr amrywiaeth o ddulliau methodolegol arloesol i sefydlu ciplun aml-lwyfan o ymddygiad ehangach y cyfrif, yn ogystal â ffocws mwy cul ar gynnwys atgas.

Ym mhennod gyntaf yr adroddiad hwn, rydym yn darparu trosolwg meintiol lefel uchel fesul llwyfan o'r dirwedd o gyfrifon sy'n berthnasol i'r DU y mae ISD wedi'u nodi fel rhai sy'n ymwneud ag ymddygiad ar-lein sy'n gysylltiedig ag eithafiaeth, terfysgaeth, iaith casineb neu gynllwyn niweidiol. Yn yr ail bennod, rydym wedi cynhyrchu map rhwydwaith o'r dirwedd gydgysylltiedig hon o weithredwyr ar-lein, gan ddefnyddio prosesu iaith naturiol i ddadansoddi negeseuon cyffredinol y cyfrifon hyn. Yn y bennod olaf, rydym yn defnyddio 'ensemble' o ddosbarthwyr iaith casineb i ddeall y naratifau casineb penodol y mae'r cymunedau hyn yn ceisio eu datblygu.

I lywio'r dadansoddiad a amlinellir yn yr adroddiad hwn, mae ISD wedi adeiladu ar ddealltwriaethau sefydledig yn y meysydd academaidd a pholisi i ddatblygu diffiniadau gweithredol (amlinellir y canfyddiadau allweddol isod) o gysyniadau allweddol - gan gynnwys 'terfysgaeth', 'eithafiaeth', 'iaith casineb', a 'cynllwyn niweidiol' - gyda'r nod o gysylltu gweithgarwch ar-lein â niwed go iawn. Roedd y diffiniadau hyn yn seiliedig yn rhannol ar droseddau blaenoriaeth a nodwyd yn y Bil Diogelwch Ar-lein[2], fel terfysgaeth, troseddau casineb, aflonyddu, bygythiadau a chymell trais. Fodd bynnag, maent yn mynd y tu hwnt i hyn ac maent hefyd yn ystyried telerau ac amodau cyfatebol rhai gwasanaethau ac arbenigedd pwnc ISD ynghylch y bygythiadau esblygol hyn.

Prif Ganfyddiadau

- **Mae'n llawer haws dod o hyd i gyfrifon sy'n gysylltiedig ag iaith casineb a chynnwys eithafol na rhai sy'n gysylltiedig â therfysgaeth ar y llwyfannau a astudiwyd ar gyfer yr adroddiad hwn.**
    - O'r 768 o gyfrifon a sianeli sy'n berthnasol i'r DU a nodwyd yn ein hastudiaeth, dim ond 55 (18 ar Instagram, 13 ar YouTube, 10 ar Facebook, 8 ar Telegram a 4 ar Twitter) oedd yn bodloni diffiniad y prosiect o derfysgaeth, gan awgrymu y gallai gweithgarwch o'r fath fod yn digwydd mewn ardaloedd mwy aneglur o'r rhyngrwyd. Roedd y rhan fwyaf o'r cyfrifon hyn (42) yn cefnogi grwpiau gwaharddedig sy'n gysylltiedig â therfysgaeth yng Ngogledd Iwerddon.
- **Canfuwyd bod y rhan fwyaf o'r cyfrifon o fewn cwmpas yr astudiaeth yn cael eu diffinio fwy gan yr amgylcheddau ar-lein atgas a chynllwynol ehangach y maent yn byw ynddynt na'u cysylltiad â grwpiau terfysgol, eithafol neu gasineb penodol.**
    - Mae ein dadansoddiad yn dangos bod gorgyffwrdd sylweddol rhwng sbectrwm eang o ddamcaniaethwyr cynllwyn a chymunedau cenedlaetholgar gwyn amlwg ar-lein. Mae twyllwybodaeth, damcaniaethau cynllwyn, casineb wedi'i dargedu, aflonyddu, ac eithafiaeth yn aml yn cydblethu ar-lein mewn ffyrdd sy'n anodd iawn eu gwahanu a'u hynysu. Canfu prosesau mapio arloesol, ar draws llwyfannau, o negeseuon o'r cyfrifon hyn, gan ddefnyddio dulliau prosesu iaith naturiol, naw 'cymuned' ieithyddol gysylltiedig, wedi'u nodweddu gan ffocws ar y cyd ar bynciau fel cynllwynion Covid-19, naratifau gwrth-mewnfudo a gwrthwynebiad i'r gymuned LGBTQ+.
- **Mae llwyfannau cyfryngau cymdeithasol mawr yn cynnwys cyfrifon sydd wedi'u codio fel rhai cas ac eithafol sy'n gallu denu cannoedd o filoedd, neu hyd yn oed filiynau, o ddefnyddwyr yn y Deyrnas Unedig.**
    - Mae gan gyfrifon a nodwyd fel rhai sy'n gysylltiedig ag eithafwyr Islamaidd nifer fawr o ddilynwyr ar Facebook – mae gan y pedwar cyfrif mwyaf dros 568,000 o ddilynwyr ar gyfartaledd.
    - Fe wnaethom nodi cyfrifon sy'n gysylltiedig ag eithafiaeth y dde eithafol, yn benodol ar Telegram (gyda rhai o sianeli mwyaf eithafiaeth y dde eithafol yn cynnwys dros 150,000 o danysgrifwyr) ac YouTube (gyda'r sianel fwyaf yn agos at 2 filiwn o danysgrifwyr). Mae'r canfyddiad hwn yn dangos bod cynnwys cas ac eithafol yn dal i gael ei rannu ar lwyfannau cyfryngau cymdeithasol mawr, er gwaethaf y ffaith bod gan ymchwilwyr dystiolaeth bod y math hwn o weithgarwch yn dod yn fwyfwy amlwg ar lwyfannau ymylol.[vii]

[2] Adeg ysgrifennu hwn, nodir y rhain yn Atodlenni 5, 6 a 7 y Bil. https://bills.parliament.uk/bills/3137

- **Rhannodd y cyfrifon yn ein hastudiaeth fwy o ddolenni i YouTube nag i unrhyw lwyfan arall.**
    - At ei gilydd, roedd cyfrifon yn ein hastudiaeth yn gysylltiedig â hyrwyddo iaith casineb, eithafiaeth, terfysgaeth neu gynnwys cynllwyn niweidiol wedi rhannu dolenni i YouTube dros 50,000 o weithiau, gan gyfrif am 78% o ddolenni i lwyfannau eraill a nodwyd yn yr astudiaeth hon. Nid oedd o fewn cwmpas yr astudiaeth benodol hon i archwilio a oedd cynnwys y dolenni hyn yn niweidiol.
    - Roedd cyfrifon sy'n gysylltiedig ag eithafwyr ar lwyfannau amrywiol fel 4chan, Twitter, Instagram a Facebook hefyd yn cyfeirio eu dilynwyr yn rheolaidd at Telegram, gan awgrymu pwysigrwydd y llwyfan yn y cymunedau hyn.
- **Er gwaethaf eu presenoldeb ar lwyfannau mwy, mae ein data'n cynnig arwyddion y gallai gweithredwyr sy'n berthnasol i'r DU, ac sy'n gysylltiedig ag iaith casineb ac eithafiaeth, fod â diddordeb mewn gwefannau llai.**
    - Roedd llwyfannau newydd fel Bitchute, Odysee, Gettr a Rumble yn gysylltiedig â'n data yn amlach na Facebook, Instagram a Reddit (ond yn llai aml na YouTube a Telegram). Mae hyn yn dangos y gallai llwyfannau o'r fath fod o ddiddordeb i gyfrifon sy'n lledaenu cynnwys sy'n gysylltiedig â therfysgaeth, eithafiaeth a chasineb.
- **Roedd cynnwys o gyfrifon yn yr astudiaeth yn cyrraedd cynulleidfaoedd sylweddol ac yn ennyn lefelau uchel o ymgysylltu ar draws llwyfannau.**
    - Roedd y cynnwys a gafodd ei uwchlwytho ar fwrdd /pol/ atgas 4chan gan ddefnyddwyr o'r Deyrnas Unedig wedi denu 1,891,328 o sylwadau. Yn ystod cyfnod yr astudiaeth, cafodd y cyfrifon a nodwyd yn ein hastudiaeth 526,398 o ymatebion ar Twitter, 462,009 o sylwadau ar Facebook, 321,830 o sylwadau ar YouTube, 179,140 o sylwadau ar Instagram, a 4,864 o sylwadau ar Reddit.
    - Lle gellir mesur hyn, gwyliwyd sianeli Telegram 95,388,986 o weithiau a fideos o gyfrifon YouTube sy'n gysylltiedig ag iaith casineb, eithafiaeth a therfysgaeth 37,429,616 o weithiau. Yn ystod cyfnod yr ymchwil, cafodd cynnwys ar y cyfrifon hyn eu hoffi 6,520,902 o weithiau ar Twitter, 3,874,941 ar Instagram a 1,569,893 ar Facebook.
- **Nid oedd llawer iawn o gynnwys a oedd yn cael ei bostio gan weithredwyr casineb ac eithafol yn cael ei ystyried yn benodol atgas gan ensemble pwrpasol o fodelau mynegi casineb.**
    - Creodd ein dull ymchwil arloesol 'ensemble' o algorithmau i adnabod iaith casineb, gan gynnwys 24 o fodelau ffynonellau agored, masnachol a phwrpasol, a 25 o eirfaoedd pwrpasol.
    - Canfu'r dull hwn (sy'n gosod trothwy uchel ar gyfer cynhwysiant, a amlinellir isod) 2,260 o negeseuon a oedd yn bodloni ein diffiniad ni o iaith casineb, a 5,371 o negeseuon a oedd yn cynnwys iaith anweddus.
    - Roedd 47% o'r enghreifftiau o iaith casineb a nodwyd yn yr astudiaeth hon yn targedu unigolion ar sail eu tarddiad cenedlaethol, 24% yn ymwneud ag iaith casineb gwrth-Foslemaidd, 15% gwrth-semitiaeth a 7% iaith casineb yn erbyn pobl Ddu.
    - Yn nodedig, mae iaith casineb amlwg yn cynrychioli cyfran fach iawn – 0.35% – o'r negeseuon cyffredinol a anfonwyd gan gyfrifon a gafodd eu cynnwys yn ein hastudiaeth.

[i]Ein Gwlad Ar-lein 2022. Ofcom. Cyrchwyd yn: https://www.ofcom.org.uk/__data/assets/pdf_file/0026/238346/ein-gwlad-ar-lein-2022.pdf

[ii] "Internet and radicalisation pathways: technological advances, relevance of mental health and role of attackers", Gweinyddiaeth Gyfiawnder y DU (2023). https://www.gov.uk/government/publications/internet-and-radicalisation-pathways-technological-advances-relevance-of-mental-health-and-role-of-attackers.

[iii] Davey, Jacob a Milo Comerford. "Between Conspiracy and Extremism: A Long COVID Threat? An Introductory Paper." *Institute for Strategic Dialogue* (2021).

Berger, J.M. "THE ALT-RIGHT TWITTER CENSUS." Vox Pol (2018). https://www.voxpol.eu/download/vox-pol_publication/AltRightTwitterCensus.pdf; Lewis, Rebecca. "Alternative influence: Broadcasting the reactionary right on YouTube." (2018); Guhl, Jakob, a Jacob Davey. "A safe space to hate: White supremacist mobilisation on telegram." Institute for Strategic Dialogue (2020); Crawford, Blyth, Florence Keen, a Guillermo Suarez-Tangil. "Memetic irony and the promotion of violence within chan cultures." (2020); Gaudette, Tiana, et al. "Upvoting extremism: Collective identity formation and the extreme right on Reddit." New Media and Society 23.12 (2021): 3491-3508; Walther, Samantha, ac Andrew McCoy. "US extremism on Telegram." Perspectives on Terrorism 15.2 (2021): 100-124.

[v] "The most popular social networks in the UK", YouGov. https://yougov.co.uk/ratings/technology/popularity/social-networks/all

[vi] Crawford, Blyth, Florence Keen, a Guillermo Suarez-Tangil. "Memetic irony and the promotion of violence within chan cultures." (2020);

[vii] Rogers, Richard. "Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media." European Journal of Communication 35.3 (2020): 213-229; Amarasingam, Amarnath, Shiraz Maher, a Charlie Winter. "How Telegram disruption impacts Jihadist platform migration." Adroddiad CREST (2021).