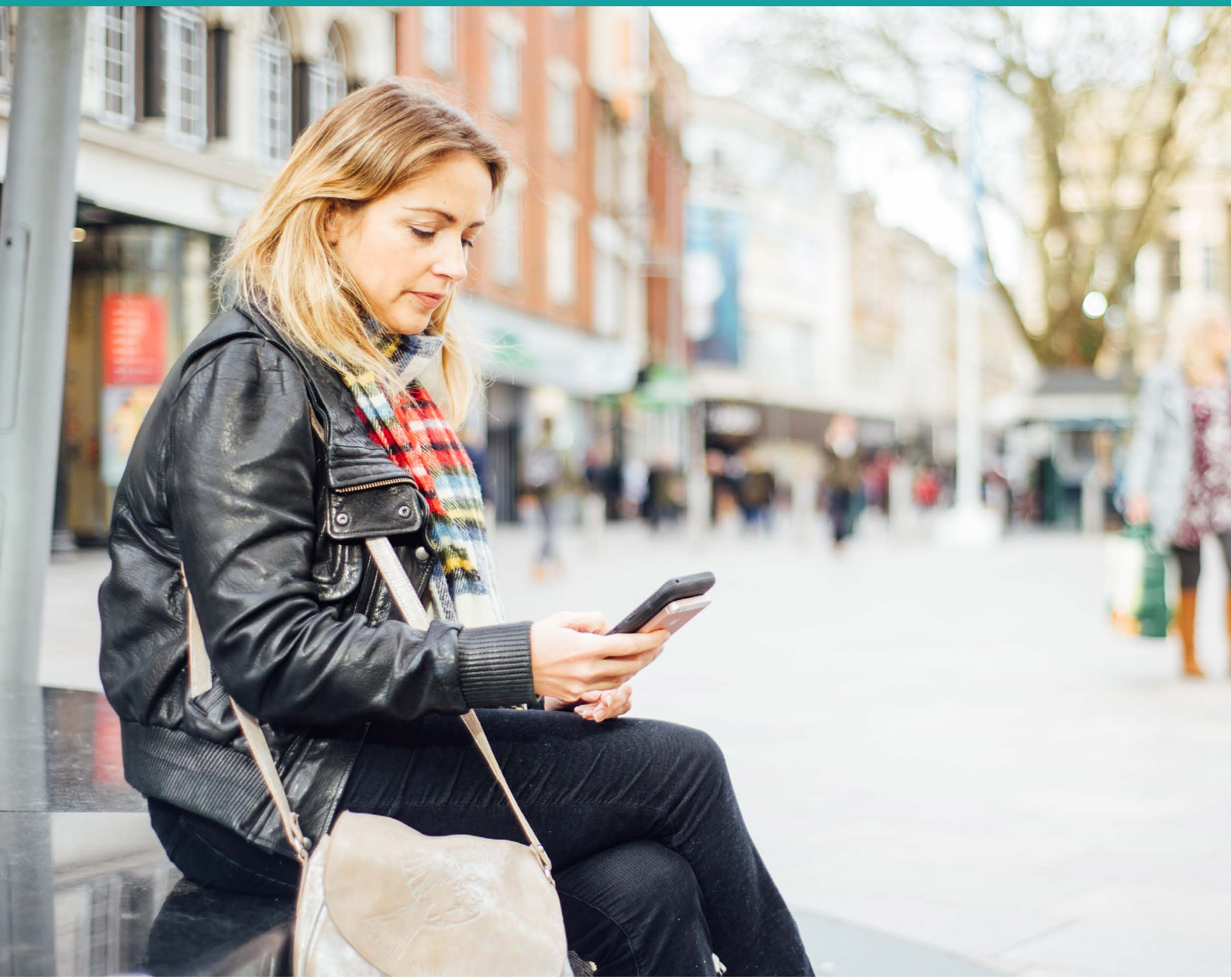
Senedd Cymru
Ymchwil y Senedd

Deallusrwydd artiffisial cynhyrchiol – trosolwg

Papur briffio

Gorffennaf 2024



www.senedd.cymru

Senedd Cymru yw'r corff sy'n cael ei ethol yn ddemocrataidd i gynrychioli buddiannau Cymru a'i phobl. Mae'r Senedd, fel y'i gelwir, yn deddfu ar gyfer Cymru, yn cytuno ar drethi yng Nghymru, ac yn dwyn Llywodraeth Cymru i gyfrif.

Gallwch weld copi electronig o'r ddogfen hon ar wefan y Senedd:
ymchwil.senedd.cymru

Gellir cael rhagor o gopïau o'r ddogfen hon mewn ffurfiau hygyrch, yn cynnwys Braille, print bras, fersiwn sain a chopïau caled gan:

Senedd Cymru
Tŷ Hywel
Bae Caerdydd
CF99 1SN

X: **@SeneddYmchwil**
Ymchwil y Senedd: **ymchwil.senedd.cymru**
Tanysgrifiwch: **Cylchlythyr**

Senedd Cymru
Ymchwil y Senedd

Deallusrwydd artiffisial cynhyrchiol – trosolwg
Papur briffio

Gorffennaf 2024

Awdur:
Amandine Debus

Mae Ymchwil y Senedd yn cydnabod y gymrodoriaeth seneddol a roddwyd i Amandine Debus gan Gyngor Ymchwil yr Amgylchedd Naturiol, a'i gwnaeth yn bosibl cwblhau'r papur briffio hwn gan y Gwasanaeth Ymchwil.

Cynnwys

Crynodeb 1

1. Beth yw deallusrwydd artiffisial cynhyrchiol a beth yw'r cyfleoedd i'r sector cyhoeddus? 3

Diffiniad 3

Perthnasedd i'r sector cyhoeddus 5

Cyfleoedd economaidd 6

Cyfleoedd penodol yng Nghymru 7

2. Beth yw'r risgiau a'r heriau? 8

Risgiau a chyfyngiadau technegol AI 8

Cyfyngiadau technegol posibl 8

Enghreifftiau o'r 'byd go iawn' 9

Risgiau y tu hwnt i gyfyngiadau technegol: pryderon rhanddeiliaid 10

Canllawiau presennol i liniaru'r risgiau er mwyn sicrhau defnydd moesegol, teg a diogel 15

3. Beth yw'r trefniadau rheoleiddio presennol ar gyfer AI cynhyrchiol yng Nghymru ac mewn gwledydd eraill? 17

Rheoliadau a chymhwysedd yng Nghymru 17

Deddfwriaeth sy'n bodoli eisoes ar gyfer Cymru 17

Cymhwysedd ar gyfer deddfwriaeth bellach 19

Polisi a deddfwriaeth y DU 20

Strategaeth ynghylch AI yn benodol 20

Goblygiadau'r Papur Gwyn ar reoleiddio AI i Gymru 22

Polisi yr UE 23

Gwledydd eraill 25

4. Beth yw'r heriau o ran cynllunio a gweithredu polisi? 28

Esblygu'n gyflym 28

Diffinio niwed ac asesu difrod 28

Yr angen am ddulliau rhyngwladol 29

Termau 30

Crynodeb

Beth yw'r cyfleoedd a'r risgiau sy'n gysylltiedig â deallusrwydd artiffisial cynhyrchiol yn y sector cyhoeddus?

Deallusrwydd artiffisial cynhyrchiol (neu AI cynhyrchiol) yw'r defnydd o ddeallusrwydd artiffisial (AI) ar gyfer creu cynnwys newydd.

Mae'n cynnig **cyfleoedd** i wella effeithlonrwydd, safon a hygyrchedd y gwasanaethau a ddarperir gan y sector cyhoeddus. Mae AI cynhyrchiol hefyd yn cynnig **cyfleoedd penodol** i hyrwyddo a chefnogi'r defnydd o'r Gymraeg.

Fodd bynnag, mae allbynnau AI cynhyrchiol yn dibynnu ar **safon y data** a ddefnyddir i hyfforddi modelau a all arwain at ragfarn, gwahaniaethu, camgymeriadau neu wybodaeth sy'n fwriadol gamarweiniol. Os oes llai o ddata ar gael i hyfforddi modelau mewn rhai ieithoedd, **fel y Gymraeg**, gall hynny hefyd gyfyngu ar allu'r AI.

Heblaw am gyfyngiadau technegol, mae'r risgiau eraill sy'n peri pryder i randdeiliaid yn cynnwys y **diffyg trefniadau rheoleiddio** presennol, yr **effeithiau ar swyddi**, **preifatrwydd**, diogelu data a **materion hawlfraint**.

Beth yw'r dull presennol o reoleiddio deallusrwydd artiffisial cynhyrchiol yng Nghymru?

Ar hyn o bryd, **nid oes dull cyfannol penodol** o ddeddfu ar gyfer AI cynhyrchiol (ac AI yn gyffredinol) yng Nghymru o ran ei ddatblygu, ei ddosbarthu neu ei ddefnyddio. Mae'r dulliau deddfwriaethol yn dibynnu ar fframwaith rheoleiddio sy'n bodoli eisoes, gan gynnwys deddfwriaeth sy'n benodol i un parth neu ddeddfwriaeth sy'n berthnasol i sawl parth.

Mae diogelu gwybodaeth bersonol, eiddo deallusol a gwasanaethau rhyngrwyd yn **faterion a gedwir yn ôl**, sy'n golygu nad oes gan Lywodraeth Cymru a'r Senedd y pŵer i ddeddfu mewn perthynas â'r meysydd hyn, a bod Cymru felly'n ddarostyngedig i'r un gyfraith â'r DU. Fodd bynnag, o ran y drefn ehangach o reoliadau diogelu data, **nid yw bob amser yn cynnwys ffin syml rhwng yr hyn 'a gedwir yn ôl' a'r hyn 'sydd wedi'i ddatganoli'.** Mae hefyd ychydig o le i'r Senedd wneud cyfraith sy'n ymwneud â **chyfle cyfartal** i fynd i'r afael ag allbynnau AI diffygiol.

Cyhoeddodd Llywodraeth y DU ei **Phapur Gwyn ar reoleiddio AI** ym mis Mawrth 2023, sy'n nodi y dylid defnyddio'r pwerau sydd eisoes gan reoleiddwyr i ymdrin ag AI, gan ddatganoli cyfrifoldeb i sectorau. Mae'n seiliedig ar ddull sy'n cefnogi arloesedd ac nid yw'n cynnwys cynllun i gyflwyno rheoliadau neu gyrff rheoleiddio penodol newydd. Wrth ymateb ym mis Chwefror 2024 **tynnodd Llywodraeth y DU** sylw at y ffaith bod rhanddeiliaid yn cefnogi'r dull, gan bwysleisio bod rheoleiddwyr eisoes yn cymryd camau i ddilyn y fframwaith arfaethedig.

Mae gwledydd eraill wedi mabwysiadu dull gwahanol yn seiliedig ar fframweithiau deddfwriaethol rhagnodol. Mae'r rhain yn cynnwys **Brasil** a **Tsieina**. Mae'r **UE** hefyd wedi bod ar flaen y gad o ran deddfwriaeth AI ac wedi datblygu **Deddf Deallusrwydd Artiffisial** (y Ddeddf AI).

Beth yw'r heriau o ran rheoleiddio AI?

Gan fod y sefyllfa o ran AI cynhyrchiol yn esblygu'n gyflym, mae **angen i'r trefniadau rheoleiddio fod yn hyblyg** i gyfrif am ddatblygiadau newydd.

At hynny, **nid yw bob amser yn hawdd** diffinio effeithiau AI cynhyrchiol, sy'n golygu y gall fod yn anodd diffinio pwy sydd wedi cael ei niweidio a phryd.

Yn olaf, **cydnabyddir bod** angen cydweithredu rhyngwladol er mwyn osgoi trosglwyddo'r problemau i leoedd eraill. Er gwaetha'r **ymdrechion** i gydweithio, er enghraifft yr Uwchgynhadledd Diogelwch AI ym mis Tachwedd 2023, mae'r strategaethau a gynlluniwyd gan wahanol wledydd yn amrywio.

1. Beth yw deallusrwydd artiffisial cynhyrchiol a beth yw'r cyfleoedd i'r sector cyhoeddus?

Diffiniad

Deallusrwydd artiffisial cynhyrchiol, y cyfeirir ato hefyd fel AI cynhyrchiol, yw'r defnydd o ddeallusrwydd artiffisial (AI) ar gyfer creu cynnwys newydd. Mae hyn yn cynnwys testun, delweddau, sain, fideos, neu gerddoriaeth. Dyma ddiffiniad yr **Alan Turing Institute** o AI cynhyrchiol:

> a type of artificial intelligence that involves creating new and original data or content. Unlike traditional AI models that rely on large datasets and algorithms to classify or predict outcomes, generative AI models are designed to learn the underlying patterns and structure of the data and generate novel outputs that mimic human creativity.

Mae AI cynhyrchiol yn dysgu patrymau o **gynnwys a grëwyd gan bobl fel tudalennau ar y we, cyfryngau cymdeithasol, a mathau eraill o gynnwys ar-lein**, ac yn aml me'n defnyddio **modelau iaith mawr ('large language models' neu 'LLM' yn Saesneg)**. Mae **modelau iaith mawr** yn dysgu o destun ac yn gallu cyflawni amrywiaeth o dasgau prosesu iaith naturiol ('natural language processing' neu 'NLP' yn Saesneg) drwy ddynwared iaith ddynol. Mae **prosesu iaith naturiol** yn disgrifio'r gangen o gyfrifiadureg sy'n canolbwyntio ar roi'r gallu i gyfrifiaduron ddeall a chynhyrchu lleferydd. Mae **modelau iaith mawr** yn dysgu cannoedd o filiynau i gannoedd o biliynau o **baramedrau** (h.y. y gwerthoedd a ddefnyddir i lunio'r model) tra byddant yn cael eu hyfforddi ac felly maent yn ddwys o safbwynt cyfrifiadurol, gallant fod yn ddrud a gallant gymryd amser hir i gael eu hyfforddi.

Yn y blynyddoedd diwethaf, mae gwelliannau o ran cofau cyfrifiadurol, maint setiau data, pŵer prosesu, a thechnegau modelu dilyniannau hir o destun wedi arwain at ddatblygu modelau iaith mawr pwerus a mwy galluog, fel OpenAI's **ChatGPT**. Mae modelau iaith mawr yn aml yn defnyddio **trawsnewidwyr**, sy'n defnyddio mecanwaith o'r enw '**hunan sylw**' i ddod o hyd i berthnasoedd a phatrymau mewn data dilyniannol.

Gellir diffinio **deallusrwydd artiffisial (AI)** fel **math o wyddoniaeth a pheirianneg sy'n creu peiriannau deallus**. Mewn geiriau eraill, mae AI yn disgrifio gallu peiriannau i **efelychu galluoedd yr ymennydd dynol** ar gyfer datrys problemau a gwneud penderfyniadau. O fewn AI, mae **dysgu peirianyddol ('machine learning' neu 'ML' yn Saesneg)** yn defnyddio data ac algorithmau i adeiladu systemau sydd â'r gallu i ddysgu. Un o is-ganghennau dysgu peirianyddol yw **dysgu dwfn ('deep**

learning' neu 'DL' yn Saesneg), sy'n ymwneud â defnyddio rhwydweithiau niwral ag iddynt haenau sy'n dysgu'n raddol, gan ddynwared sut y mae niwronau yn gweithio.

Mae'r **rhan fwyaf o dechnolegau AI cynhyrchiol diweddar** yn **dibynnu ar ddefnyddio modelau sylfaen**, sef modelau dysgu peirianyddol ar raddfa fawr sydd wedi'u hyfforddi gan ddefnyddio swm helaeth o ddata. Gall modelau sylfaen wneud mwy na **chyflawni tasgau penodol yn unig** a gellir eu haddasu i amrywiaeth o barthau. Yn aml, fe'i gelwir yn fodelau **at ddefnydd cyffredinol**. Fodd bynnag **gall AI cynhyrchiol** gael eu cynllunio ar gyfer tasgau penodol cul a defnyddio dulliau nad ydynt yn fodelau sylfaen (e.e. mae **technoleg ffugio dwfn** ('deepfake' yn Saesneg), a ddiffinnir yn Nodyn 1 isod, wedi cael ei defnyddio ers 2014 ac yn defnyddio math arall o rwydweithiau niwral).

Nodyn 1

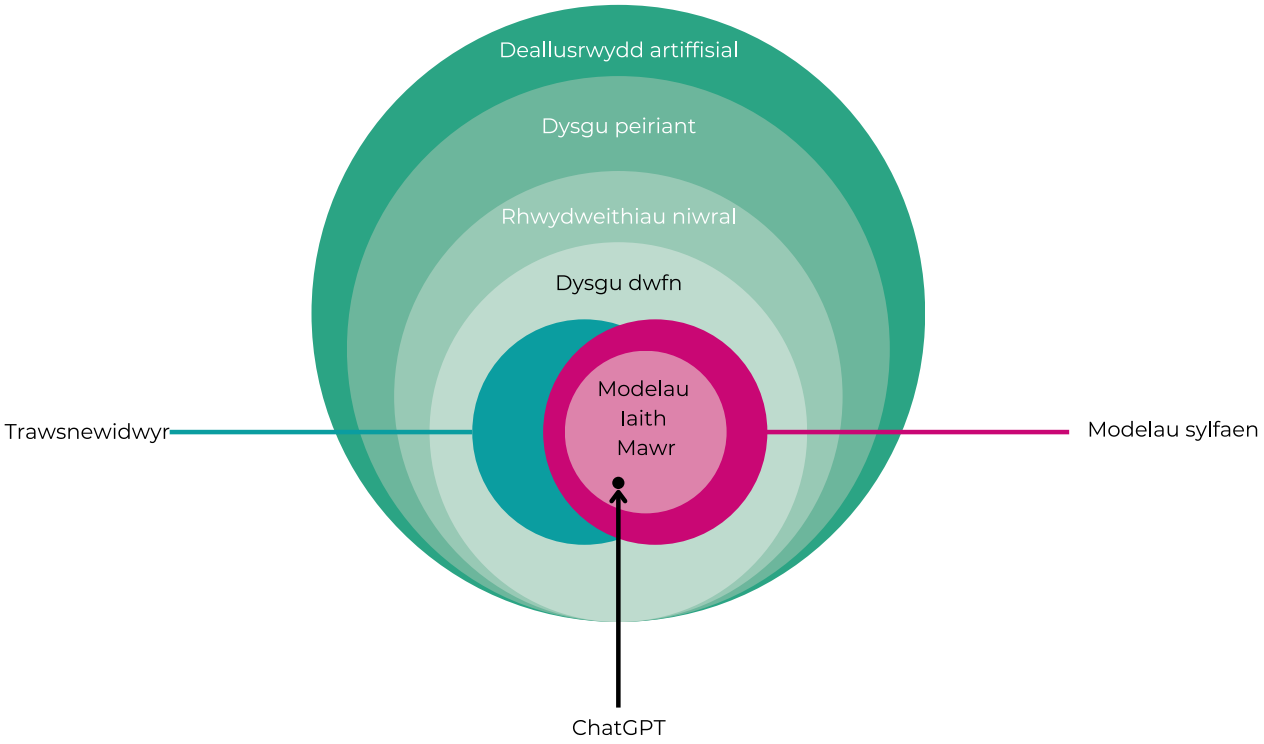
Mae technoleg **ffugio dwfn** yn disgrifio cyfryngau synthetig sy'n disodli wyneb neu lais un person ac yn rhoi wyneb neu lais rhywun arall yn eu lle, mewn ffordd sy'n edrych yn wir (mewn fideos neu recordiadau sain yn aml, ond gallai ddigwydd mewn lluniau hefyd).

Gellir defnyddio technegau amrywiol i hyfforddi modelau AI cynhyrchiol, er enghraifft:

- **dysgu dan oruchwyliaeth** h.y. mae'r model yn cael ei hyfforddi gan ddefnyddio **data sydd wedi'u labelu** (h.y. data sydd wedi'u tagio i nodi eu priodweddau yng nghyd-destun y dasg benodol a gyflawnir gan y model AI);
- **dysgu heb oruchwyliaeth** h.y. nid yw'r data hyfforddi wedi'u labelu;
- **dysgu dan oruchwyliaeth rannol**, h.y. mae rhan fach o'r data hyfforddi wedi'i labelu a rhan fawr heb ei labelu;
- **dysgu hunan oruchwylio**, h.y. nid yw'r data hyfforddi wedi'u labelu ac mae'r model yn darogan ffug-labeli iterus ar eu cyfer; a
- **atgyfnerthu'r hyn a ddysgir**, h.y. mae'r model yn dysgu drwy ddull mentro a methu ac yn dysgu o'i gamgymeriadau.

Gellir **cyfuno** gwahanol fathau o ddysgu. Er enghraifft, gellir hyfforddi modelau cynhyrchiol gan ddefnyddio dysgu heb oruchwyliaeth, ac yna eu hyfforddi ymhellach gan ddefnyddio dysgu dan oruchwyliaeth ar gyfer tasg benodol.

Ffigur 1: diagram Venn yn disgrifio'r gwahaniaethau cyffredin rhwng mathau o AI



Defnyddir **termau eraill** weithiau, fel 'modelau ffin eithaf' i ddisgrifio modelau sydd ar flaen y gad, neu 'deallusrwydd cyffredinol artiffisial (AGI)' ac 'AI cryf' i ddisgrifio AI all gyflawni unrhyw dasg y gallai dyn ei chyflawni. Fodd bynnag, mae'r termau hyn yn aml yn cael eu herio oherwydd anghysondeb y diffiniadau. O ran modelau 'ffin eithaf', nid oes ffordd gyson o fesur 'eithafedd y ffin', ac nid yw AGI/AI 'cryf' yn disgrifio galluoedd AI cyfredol.

Perthnasedd i'r sector cyhoeddus

Mae nifer o gwmnïau technoleg ac ymgynghori, fel **Microsoft**, **Deloitte** a **Google**, wedi tynnu sylw at y cyfleoedd y mae AI cynhyrchiol yn eu cynnig i'r sector cyhoeddus. Mae **athrawon** a nifer o **gyrff cyhoeddus** fel yr **Awdurdod Ymddygiad Ariannol**, yr **Asiantaeth Rheoleiddio Meddyginiaethau a Chynhyrchion Gofal Iechyd** a'r **Swyddfa Gyfathrebu (Ofcom)** hefyd wedi pwysleisio manteision AI cynhyrchiol.

Mae AI cynhyrchiol yn cynnig y cyfleoedd a ganlyn:

- **gwella effeithlonrwydd**;
- **gwella boddhad gweithwyr** drwy leihau tasgau diflas;

- **gwella gwasanaethau dinasyddion (e.e. drwy ddefnyddio sgwrsfotiau)**;
- **darparu cymorth creadigol**;
- **dadansoddi data dwfn** ar gyfer allbynnau cynhwysfawr; a
- **gwella hygyrchedd a chynhwysiant**.

Hefyd, mae llawer o **gyfleoedd penodol i seneddau**, er enghraifft argymhellion ar gyfer deddfwriaeth sy'n seiliedig ar fylchau a nodwyd, creu crynodebau yn dilyn ymgyngoriadau, neu sganio'r gorwel ar gyfer ymchwil wyddonol. Nid yw'r rhestr hon yn un gynhwysfawr a dylid ystyried **manylion y sefyllfa** ym mhob lleoliad seneddol a'u strategaethau digidol i asesu perthnasedd a blaenoriaeth.

Mae AI cynhyrchiol eisoes wedi cael ei ddefnyddio yn y sector cyhoeddus ledled y byd. Er enghraifft, fe'i defnyddiwyd yn y **sector addysg** gan athrawon i gynhyrchu adnoddau addysgol wedi'u teilwra ar gyfer myfyrwyr unigol neu i greu adborth ar waith myfyrwyr. Mae myfyrwyr wedi ei ddefnyddio, er enghraifft, i'w helpu i gwblhau aseiniadau, i egluro cysyniadau neu ddod o hyd i wybodaeth, ac i gynhyrchu cynnwys ar gyfer cyflwyniadau. Mae hefyd wedi cael **ei ddefnyddio gan gynghorau** i greu sgwrsfotiau i ymateb i ymholiadau. Mae **cynlluniau peilot wedi eu cynnal** ar gyfer drafftio deddfwriaethol ym Mrasil, yr Ariannin, y Ffindir a'r Eidal. Ym **Mrasil**, gall y cyhoedd ofyn am allbynnau seneddol (e.e. Biliau) a defnyddir AI cynhyrchiol i dynnu gwybodaeth o'r rheini.

Yng Nghymru, mae AI cynhyrchiol **wedi cael ei ddefnyddio** gan staff academaidd a myfyrwyr, ac mae Prifysgol De Cymru wedi cyhoeddi **canllawiau** ar sut i'w ddefnyddio'n briodol. Defnyddiwyd ChatGPT hefyd i ysgrifennu **araith** a draddodwyd yn Senedd Cymru.

Cyfleoedd economaidd

Yn ôl gwerthusiad **McKinsey**, gallai effaith AI cynhyrchiol ar gynhyrchiant ychwanegu triliynau o ddoleri mewn gwerth i'r economi fyd-eang. Mae **KPMG** yn amcangyfrif y gallai AI cynhyrchiol ychwanegu £31 biliwn o ran cynnyrch domestig gros i economi'r DU yn y degawd nesaf. Yn gyffredinol, gallai AI arwain at gynnydd o £7.9 biliwn (twf o 9.8 y cant o ran cynnyrch domestig gros) yng Nghymru erbyn 2030, yn ôl adroddiad a gyhoeddwyd yn 2017 gan **PwC**.

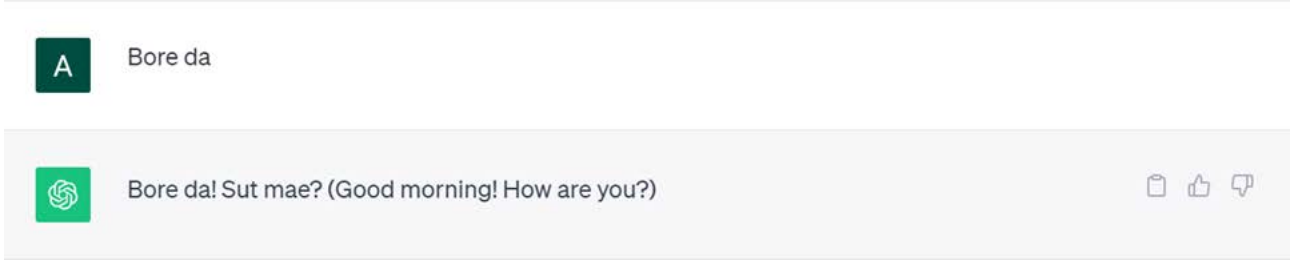
Fodd bynnag, mae **ansicrwydd** arwyddocaol o hyd ynghylch effeithiau economaidd byd-eang AI cynhyrchiol.

Cyfleoedd penodol yng Nghymru

Gan fod Cymru yn wlad ddwyieithog, mae AI cynhyrchiol yn cynnig cyfleoedd penodol i hyrwyddo a chefnogi'r defnydd o'r Gymraeg. Mae'r rhan fwyaf o'r sector cyfieithu yng Nghymru, gan gynnwys Senedd Cymru, **eisoes yn defnyddio** model cyfieithu peirianyddol niwral ar gyfer y Gymraeg, sy'n **darogan** dilyniant tebygol o eiriau yn Gymraeg, ar ôl cael dilyniant o eiriau yn Saesneg. Mae adnoddau adnabod lleferydd awtomatig Cymraeg 'AI traddodiadol' (h.y. AI nad yw'n gynhyrchiol) wedi cael eu datblygu hefyd, fel y **cynorthwyydd llais Macsen**. Gellid gwella'r rhain gydag AI cynhyrchiol yn y dyfodol, yn dilyn **esiampl Alexa**.

Mae **Cynllun Gweithredu Technoleg Cymraeg** Llywodraeth Cymru, a gyhoeddwyd yn 2018, yn tynnu sylw at bwysigrwydd sgwrsfotiau Cymraeg fel y gall dinasyddion wneud ymholiadau i'w cynghorau lleol yn Gymraeg. Mae'r Cynllun Gweithredu hwn yn cynnwys pecyn gwaith sy'n benodol ar gyfer dysgu peirianyddol a sgwrsio gan AI drwy gyfrwng y Gymraeg.

Ffigur 2: Dechrau sgwrs yn Gymraeg ar ChatGPT-3.5



Ddiwedd 2023, lansiodd y **Ganolfan Gwasanaethau Cyhoeddus Digidol** arolwg i gael rhagor o wybodaeth am sut y mae'r sector cyhoeddus yng Nghymru yn defnyddio AI, gan gynnal cyfweliadau â darparwyr gwasanaethau cyhoeddus. **Cyhoeddodd y Ganolfan adroddiad** ddechrau 2024 yn argymell creu **cymuned ymarfer ar gyfer awtomeiddio ac AI** i sefydliadau'r sector cyhoeddus yng Nghymru.

2. Beth yw'r risgiau a'r heriau?

Risgiau a chyfyngiadau technegol AI

Cyfyngiadau technegol posibl

Mae allbynnau AI cynhyrchiol yn dibynnu ar **safon y data hyfforddi** a gaiff ei fwydo i mewn i'r algorithmau. Mae algorithm yn **cyfateb** i'r cyfarwyddiadau a gaiff peiriant er mwyn dod o hyd i atebion i gwestiwn neu broblem. O ganlyniad, mae'r risgiau posibl o ran allbynnau yn cynnwys:

- **rhagfarn a diffyg niwtraliaeth**;
- **monoddiwylliant** a **rhwystrau rhag datblygu safbwyntiau lluosog** (esboniad pellach yn Nodyn 2 isod);
- **gwybodaeth su'n fwriadol gamarweiniol (twyllwybodaeth)**;
- **camgymeriadau, creu gwybodaeth anwir** a **chamwybodaeth**; a
- **gwahaniaethu**.

Nodyn 2

Bydd allbynnau AI cynhyrchiol yn **adlewyrchu gwerthoedd** crewyr y data hyfforddi, neu'r datblygwyr a ddewisodd y setiau data hyfforddi i'w defnyddio. O ganlyniad, yn aml, ni fydd lleisiau ymylol yn cael eu cynrychioli mewn allbynnau AI cynhyrchiol.

Gall hyn arwain at **effeithiau ehangach** ar y gymdeithas gyfan. Er enghraifft, gall arwain at ddirywiad pellach o ran ffydd pobl mewn gwybodaeth, gall ddylanwadu ar wleidyddiaeth a chymdeithas, ac arwain at ailadrodd methiannau os defnyddir allbynnau o'r fath mewn sefydliadau ar raddfa fawr.

Gallai'r broblem hon **waethygu** wrth ddatblygu modelau i'r dyfodol, sy'n debygol o ddefnyddio allbynnau modelau blaenorol fel data hyfforddi, gyda'r risg felly y cânt eu cyfaddawdu oherwydd am fod allbynnau gwallus wrth wraidd iddynt.

At hynny, mae'n aml yn anodd gwybod pam y mae gwall wedi digwydd, ac mae'n anodd bwydo gwybodaeth ddynol i mewn i'r cylch am fod diffyg goruchwyliaeth o'r ffordd y mae'r algorithm yn gweithio. Dyna pam y mae llawer o systemau AI yn cael eu galw'n '**flychau du**': mae gan y datblygwyr ddealltwriaeth gyfyngedig o'r mecanweithiau mewnol. Mae'r **Alan Turing Institute** wedi tynnu sylw at y problemau hyn o ran dehongli, ac mae **academyddion** wedi galw am fodelau y

gellir eu dehongli.

O ran Cymru yn benodol, un cyfyngiad posibl, o leiaf ar hyn o bryd, yw'r **diffyg data hyfforddi yn Gymraeg**, sy'n cyfyngu ar allu AI cynhyrchiol i ddarparu allbynnau boddhaol. Yn fwy cyffredinol, mae AI cynhyrchiol **yn aml yn cael ei hyfforddi** ar ddata o'r rhyngrwyd, sydd ar gael yn bennaf mewn nifer fach o ieithoedd sy'n cynnig llawer o ddata (e.e. Saesneg, Sbaeneg, Mandarin), gan greu bylchau o ran unrhyw iaith nad oes llawer o ddata ar gael ynddi. Gallai'r allbwn a gynhyrchir gan declyn AI cynhyrchiol ar gyfer cais Saesneg amrywio'n fawr o gymharu ag allbwn ar gyfer cais Cymraeg. Gallai fod goblygiadau hefyd o ran rhagfarn wrth ddefnyddio ieithoedd gwahanol.

Yn olaf, dangosodd **astudiaeth** y gall ymddygiad systemau AI cynhyrchiol newid yn gymharol gyflym dros amser, drwy ddangos gostyngiad yng ngallu ChatGPT i ddilyn cyfarwyddiadau defnyddwyr dros dri mis, gan ddangos yr angen am fonitro parhaus.

Enghreifftiau o'r 'byd go iawn'

Mae nifer o **brofion arbrofol** wedi dangos cyfyngiadau technegol AI cynhyrchiol ar gyfer cymwysiadau yn y 'byd go iawn', hyd yn oed ar gyfer tasgau sylfaenol y gallai bodau dynol nad ydynt yn arbenigwyr eu datrys.

Mae AI cynhyrchiol hefyd wedi cynhyrchu allbynnau sy'n gwaethygu stereoteipiau niweidiol **sawl gwaith**, neu wneud datganiadau rhywiaethol a hiliol, **hyd yn oed pan fydd yr allbynnau hynny'n cael eu hidlo i gael gwared ar sylwadau rhagfarnol**.

Yn gyffredinol, gwelwyd problemau o ran rhagfarn wrth ddefnyddio cymwysiadau AI y 'byd go iawn' yn y sector cyhoeddus. Er enghraifft, dyfarnwyd bod technoleg adnabod wynebau a ddefnyddiwyd gan Heddlu De Cymru yn **anghyfreithlon**, gan ei bod yn torri hawliau dynol oherwydd diffyg arweiniad ar sut i'w defnyddio; diffyg ystyriaeth o'r effaith o ran diogelu data; a methiant i fynd i'r afael yn ddigonol â risgiau o ran rhagfarn hiliol a rhywiol. Fodd bynnag, pe na bai'r materion hyn wedi bod yn rhan o'r achos, dywedodd y Llys Apêl y byddai wedi bod yn gyfreithlon i ddefnyddio technoleg adnabod wynebau. Mae enghreifftiau eraill y tu allan i'r DU yn cynnwys **algorithm darogan twyll** a ddefnyddir gan Awdurdodau Treth yr Iseldiroedd y canfuwyd ei fod yn gwahaniaethu'n anuniongyrchol ar sail hil ac ethnigrwydd. Cafodd rhieni eu cyhuddo o dwyll ar gam, gan arwain at ddiweithdra, methdaliadau ac ysgariadau. Ymddiswyddodd y Llywodraeth yn y pen draw a thalwyd €5 biliwn mewn iawndal.

Risgiau y tu hwnt i gyfyngiadau technegol: pryderon rhanddeiliaid

Heblaw am gyfyngiadau technegol AI cynhyrchiol oherwydd y data a gaiff eu mewnbynnu, mae risgiau eraill yn ymwneud â rhanddeiliaid gan gynnwys:

1. **Diffyg rheoleiddio**: Mae arbenigwyr AI cynhyrchiol, **fel cyd-sylfaenydd OpenAI**, wedi gwthio am ddeddfwriaeth i reoleiddio defnydd a datblygiad AI. Mae'r **Ada Lovelace Institute** yn argymell bod angen fframwaith rheoleiddio mwy cyfannol a chynhwysfawr ar gyfer AI yn y DU, a bod angen eglurder o ran atebolrwydd yn enwedig, ymhlith pethau eraill. Yn ôl **arolwg** a gynhaliwyd gan yr Ada Lovelace Institute, hoffai 62 y cant o bobl Prydain weld cyfreithiau a rheoliadau ynghylch defnyddio technolegau AI. Galwodd y grŵp eiriolaeth defnyddwyr BEUC am **fwy o reoleiddio**, yn bennaf oherwydd pryderon preifatrwydd (mae rhagor o fanylion am bryderon preifatrwydd yn Rhan 5 isod). Mae Cyngres yr Undebau Llafur (TUC) hefyd **wedi galw** am ddeddfwriaeth i ddiogelu hawliau gweithwyr ac wedi lansio tasglu AI. Cyhoeddodd y tasglu **Fil AI a Chyflogaeth drafft** ym mis Ebrill 2024 ac mae'n **lobïo** iddo gael ei ymgorffori yng nghyfraith y DU.
2. **Effeithiau ar swyddi a'r posibilrwydd o golli swyddi**: Mae **KPMG** yn amcangyfrif y bydd AI cynhyrchiol yn effeithio ar 40 y cant o swyddi'r DU. Mae **gweithwyr creadigol** wedi mynegi pryderon. Yn ogystal ag awduron a chyfieithwyr, mae **KPMG** hefyd yn rhagweld effeithiau ar raglenwyr cyfrifiadurol a gweithwyr datblygu meddalwedd proffesiynol oherwydd gallu AI cynhyrchiol i ysgrifennu cod. Mae Adran Addysg y DU wedi llunio **adroddiad ar effaith AI ar swyddi a hyfforddiant yn y DU**, a dynnodd sylw at effaith sylweddol modelau iaith mawr ar alwedigaethau proffesiynol y mae angen gradd i'w cyflawni fel arfer, yn enwedig y rhai sy'n cynnwys gwaith clerigol mewn rolau cyllid, y gyfraith a rheoli busnes. Fodd bynnag, nododd yr adroddiad hefyd mai Cymru yw un o ranbarthau'r DU y mae disgwyl i AI effeithio leiaf ar weithwyr yno yn gyffredinol. **Lluniwyd papur** gan yr Institute for the Future of Work yn seiliedig ar arolwg o dros fil o gwmnïau'r DU i ddeall y broses o fabwysiadu AI a'r effeithiau posibl ar swyddi. Daeth i'r casgliad bod y canlyniadau'n ansicr, ond bod angen gweithredu ar frys i osgoi sefyllfa lle byddai anghydraddoldebau rhanbarthol a demograffig presennol yn gwaethygu. Cytunodd **Cyngor Partneriaeth y Gweithlu** i gyhoeddi papur ar **gyfleoedd a bygythiadau i'r sector cyhoeddus yn deillio o ddeallusrwydd artiffisial**, a dylai hyn ddarparu rhagor o wybodaeth am yr effeithiau ar swyddi yn y sector cyhoeddus yn benodol.

   Ym mis Awst 2023, aeth TUC Cymru ati, gyda Kings College Llundain a

Connected by Data, i **ddechrau ymchwiliad** i ymateb gweithwyr i'r defnydd o AI ar draws sectorau. Mae'r ymchwiliad hwn wedi tynnu sylw at y ffaith bod gweithwyr yn poeni y gallai AI, a digideiddio yn gyffredinol, waethygu'r sefyllfa yn y gweithle. Er enghraifft, dangosodd yr ymchwiliad hwn bryderon ynghylch **gwyliadwriaeth gan reolwyr**, **diswyddiadau a diffyg ymgynghori gydag undebau** ac **amheuaeth** ynghylch galluoedd AI o'i gymharu â sgiliau dynol. **Lluniodd y TUC adroddiad** yn dangos cynnydd arwyddocaol o ran gwyliadwraeth ar weithwyr, a **chanllawiau** ar gyfer undebau ar AI a'r gweithle. At hynny, tynnodd y TUC sylw at y risg o ragfarn a gwahaniaethu sy'n **dechrau ar y cam recriwtio**. Cyhoeddodd Llyfrgell Tŷ'r Cyffredin bapur ar **ddeallusrwydd artiffisial a chyfraith cyflogaeth** a dynnodd sylw at risgiau tebyg o ddefnyddio AI ar gyfer recriwtio, rheoli llinell, a gwyliadwriaeth.

3. **Adnabod allbynnau a gynhyrchwyd gan gyfrifiadur**: Tynnodd Llywodraeth y DU sylw at y ffaith y **gall fod yn anodd, gyda chyfryngau synthetig a gynhyrchir gan AI, adnabod mai dyna beth ydynt**, a bod y dulliau dilysu sydd ar gael yn annibynadwy am y tro. Nodwyd y mater hwn hefyd gan yr **Alan Turing Institute**. **Pwysleisiodd yr Institute for Government** y broblem o ran technoleg ffugio dwfn a ffydd wrth ryngweithio, a allai arwain at benderfyniadau gan lywodraeth.

Ffigur 3: Delwedd a gynhyrchwyd gan AI ar ôl rhoi 'Cymru' fel cais ar Pixlr

4. **Technolegau sy'n esblygu'n gyflym**: Pwysleisiodd yr Alan Turing Institute bod AI cynhyrchiol yn esblygu yn hynod o gyflym, sy'n golygu y gall risgiau newydd ddod i'r amlwg yn gyflym.

5. **Seiberddiogelwch a phreifatrwydd**: Yn adroddiad **Llywodraeth y DU ar y risgiau diogelwch a berir gan ddeallusrwydd artiffisial cynhyrchiol hyd at 2025**, nodir y gall AI cynhyrchiol helpu i awtomeiddio ymosodiadau seiber a chynyddu gwendidau digidol drwy:

> (...) corrupting training data ('data poisoning'), hijacking model output ('prompt injection'), extracting sensitive training data ('model inversion'), misclassifying information ('perturbation') and targeting computing power.

Gallai ymholiadau sy'n cael eu storio ar-lein hefyd ei gwneud yn haws i hacwyr **gael gafael ar ddata a allai fod yn sensitif**. Mae pryderon preifatrwydd wedi arwain at **wahardd ChatGPT yn yr Eidal dros dro yn 2023**. Cafodd ChatGPT ei **adfer** fis yn ddiweddarach ar ôl 'gwella tryloywder a hawliau i ddefnyddwyr Ewropeaidd', gan gynnwys opsiwn i beidio â defnyddio sgyrsiau ar gyfer algorithmau hyfforddi ChatGPT, mesurau gwirio oedran, a rhybudd yn egluro y gallai ChatGPT ddarparu gwybodaeth anghywir.

6. **Hawlfraint a defnydd teg, eiddo deallusol, llên-ladrad**: Mae AI cynhyrchiol yn tynnu data o wahanol ffynonellau, ac yn **crynhoi neu'n dynwared cynnwys sy'n bodoli eisoes, yn aml heb ganiatâd y perchnogion**. Argymhellodd Swyddfa Eiddo Deallusol y DG fod angen egluro'r sefyllfa o ran eiddo deallusol ac AI cynhyrchiol, ac mae'n gweithio ar god ymarfer ar hawlfraint ac AI. **Cyflwynwyd achosion cyfreithiol yn UDA** yn ymwneud â materion hawlfraint gyda chod a chelf weledol a gynhyrchir o AI cynhyrchiol. **Mae'r Alan Turing Institute** wedi tynnu sylw at faterion llên-ladrad yng nghyd-destun addysg ac asesiadau.
7. **Camfanteisio a chostau dynol**: Yn y broses o ddatblygu AI cynhyrchiol, mae risg o **fanteisio ar weithwyr**. Caiff modelau eu gwirio yn aml drwy broses o **atgyfnerthu'r hyn a ddysgir drwy adborth dynol ('reinforcement learning from human feedback' neu 'RLHF' yn Saesneg)**, sy'n cynnwys adolygwyr dynol. Yn ôl **adroddiad gan Goleg Prifysgol Llundain**, mae adolygwyr RLHF ChatGPT yn dod o dde'r byd yn bennaf ac yn cael incwm isel am eu gwaith (e.e. llai na $3 yr awr yn Kenya), a hynny wrth ymdrin â chynnwys sy'n effeithio'n niweidiol ar eu llesiant.
8. **Effaith amgylcheddol**: Mae hyfforddi AI cynhyrchiol, yn enwedig AI at ddibenion cyffredinol, yn waith sy'n **defnyddio pŵer sydd ag ôl troed carbon sylweddol**. Yn ôl adroddiad gan Goleg Prifysgol Llundain:

   > (...) it is estimated that the training of GPT3 (the GPT used by the first version of ChatGPT made available to the public) consumed 1,287 megawatt hours of electricity and generated 552 tons of carbon dioxide, the equivalent of 123 cars driven for one year.

9. **Colli rheolaeth a dibyniaeth**: Mae colli rheolaeth dros ddata sy'n cael eu bwydo i mewn i fodelau yn risg a grybwyllwyd gan **PwC**. Gall gorddibyniaeth ar AI cynhyrchiol hefyd arwain at oddefoldeb wrth ddysgu, diffyg ymgysylltu, ac effeithio ar ddatblygu sgiliau fel gwaith tîm, cyfathrebu a meddwl yn feirniadol, **yn ôl yr Alan Turing Institute**.
10. **Dibyniaeth fasnachol**: **Mae'r Alan Turing Institute** wedi rhybuddio yn erbyn dibyniaeth ar AI cynhyrchiol masnachol, a allai arwain at ddiffyg tryloywder a diffyg eglurder o ran atebolrwydd. Mae'r **Comisiwn Cydraddoldeb a Hawliau**

Dynol wedi pwysleisio hefyd nad yw nifer o gyfranogwyr AI preifat yn rhannu eu technoleg oherwydd sensitifrwydd masnachol.

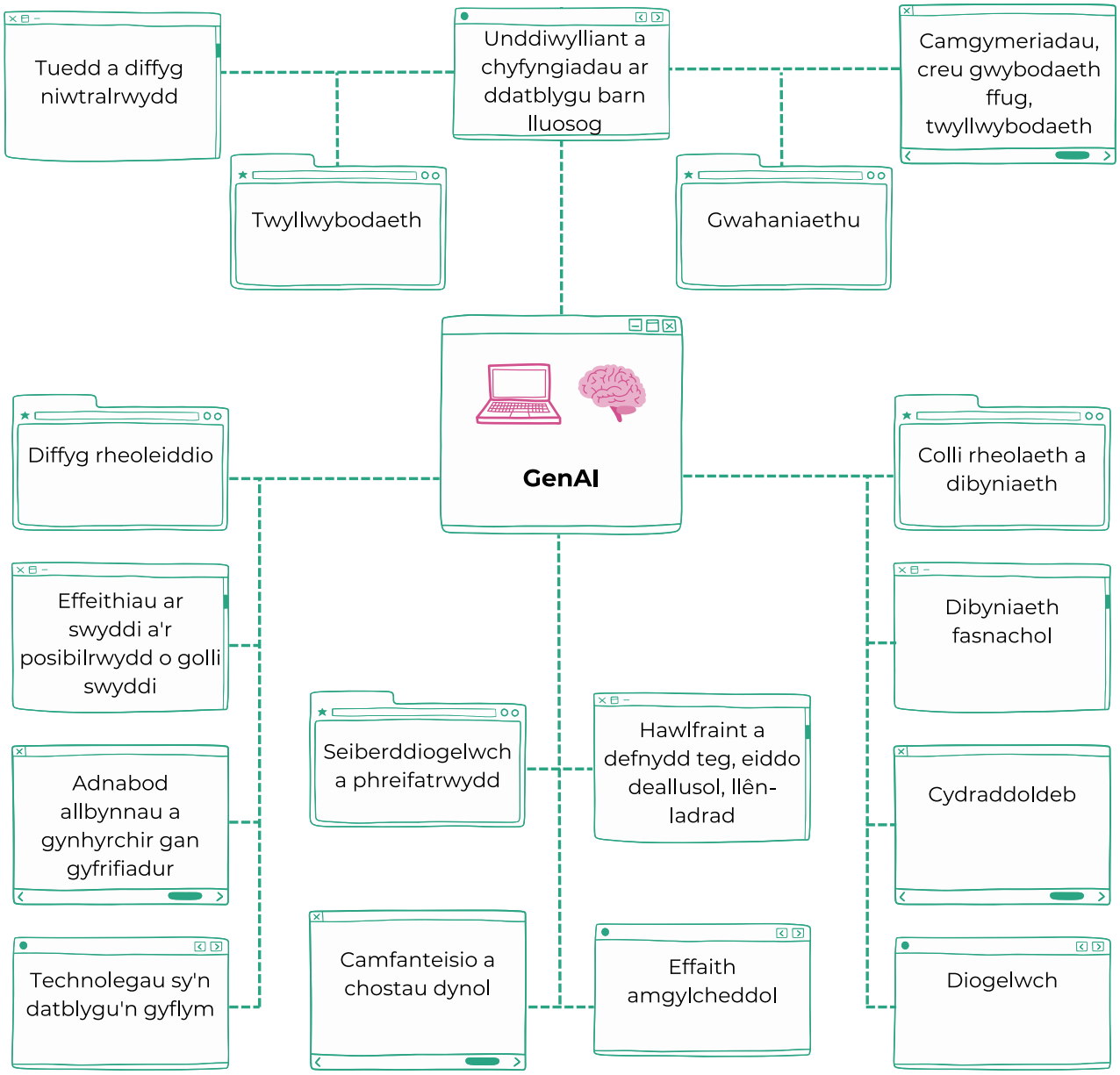
11. **Cydraddoldeb**: **Nododd y Comisiwn Cydraddoldeb a Hawliau Dynol** ei bod yn ymddangos nad yw'r rhan fwyaf o awdurdodau lleol yn ystyried cydraddoldeb fel rhan o'u prosesau caffael, ac nad yw rhai ohonynt yn cydnabod neu'n rhoi gwybod eu bod yn defnyddio AI cynhyrchiol. Nododd y Comisiwn fod dryswch ar hyn o bryd rhwng moeseg a chydraddoldeb, gan gyfeirio at yr **Asesiad o'r Effaith ar Ddiogelu Data (DPIA)** fel enghraifft o hyn. Proses i helpu pobl i nodi a lleihau risgiau diogelu data sydd ynghlwm wrth brosiect yw'r DPIA. Mae'n orfodol ar gyfer unrhyw brosesu sy'n debygol o arwain at risg uchel i unigolion ac ystyrir mai dyma'r arfer gorau ar gyfer unrhyw brosiect sy'n trin data personol. Dywedodd y Comisiwn Cydraddoldeb a Hawliau Dynol fod awdurdodau lleol yn aml yn credu eu bod yn mynd i'r afael â materion cydraddoldeb gyda DPIA, ond yn aml nid oedd hynny'n wir a bod angen gwneud mwy. Dywedodd y Comisiwn hefyd nad oedd y **Papur Gwyn ar AI** a gyhoeddwyd gan Lywodraeth y DU (rhagor o fanylion yn rhan 3) yn pwysleisio'r fframwaith cydraddoldeb a hawliau dynol ddigon. Mae'r **Alan Turing Institute** hefyd yn rhybuddio yn erbyn y posibilrwydd o ehangu'r ffin ddigidol a thorri hawliau plant (e.e. camfanteisio masnachol).

12. **Diogelwch**: Yn adroddiad **Llywodraeth y DU ar y risgiau diogelwch a berir gan ddeallusrwydd artiffisial cynhyrchiol hyd at 2025** nodir fel a ganlyn:

> Generative AI can be used to assemble knowledge on physical attacks by non-state violent actors, including for chemical, biological and radiological weapons.

Mae'r adroddiad hefyd yn sôn nad yw mesurau diogelwch bob amser yn effeithiol, ac y bydd AI cynhyrchiol yn helpu i gyflymu'r broses o gael gwared ar rwystrau cyfredol fel mynediad at gydrannau neu wybodaeth ddealledig.

Ffigur 4: Crynodeb o risgiau AI cynhyrchiol



Canllawiau presennol i liniaru'r risgiau er mwyn sicrhau defnydd moesegol, teg a diogel

Mae amryw o ganllawiau anstatudol yn bodoli i liniaru'r risgiau a restrir uchod sy'n gysylltiedig ag AI cynhyrchiol. Dyma'r canllawiau sy'n berthnasol i'r sector cyhoeddus yng Nghymru:

- Mae canllaw yr Alan Turing Institute, '**Guide for the responsible design and implementation of AI systems in the public sector**' (2019), yn trafod pob math o AI, nid dim ond AI cynhyrchiol;
- Mae cyhoeddiad y Swyddfa Digidol a Data Ganolog a'r Swyddfa Deallusrwydd Artiffisial, '**Guide to using artificial intelligence in the public sector**' (2019), yn

cynnwys **canllawiau o ran moeseg a diogelwch**: mae'n trafod pob math o AI;

- Nid yw cyhoeddiad y Swyddfa Digidol a Data Ganolog, '**Data Ethics Framework**' (2020), yn ymwneud yn benodol ag AI neu AI cynhyrchiol;
- Canllaw y Comisiwn Cydraddoldeb a Hawliau Dynol, '**Public Sector Equality Duty (PSED) Artificial Intelligence Guidance**' (2022): mae'n werth nodi, er bod y canllawiau hyn yn berthnasol i Gymru, **nad oedd yr asesiad trylwyr a gynhaliwyd gan y Comisiwn o ran Dyletswydd Cydraddoldeb y Sector Cyhoeddus ac AI yn cynnwys Cymru**;
- Mae adroddiad y Swyddfa Digidol a Data Ganolog a'r Ganolfan Moeseg Data ac Arloesi '**Transparency Recording Standard - Guidance for Public Sector Bodies**' (Ionawr 2023), yn berthnasol i unrhyw offeryn algorithmig, nid dim ond AI cynhyrchiol;
- Canllaw Swyddfa'r Cabinet a'r Swyddfa Digidol a Data Ganolog, '**Guidance to civil servants on use of generative AI**' (Medi 2023);
- Papur polisi'r Adran Addysg, '**Generative artificial intelligence (AI) in education**' (Hydref 2023); a
- Canllawiau'r Ganolfan Seiberddiogelwch Genedlaethol '**Guidelines for secure AI system development**' (Tachwedd 2023).

3. Beth yw'r trefniadau rheoleiddio presennol ar gyfer AI cynhyrchiol yng Nghymru ac mewn gwledydd eraill?

Rheoliadau a chymhwysedd yng Nghymru

Deddfwriaeth sy'n bodoli eisoes ar gyfer Cymru

Ar hyn o bryd, **nid oes dull cyfannol penodol** o ddeddfu ar gyfer AI cynhyrchiol (ac AI yn gyffredinol) yng Nghymru o ran ei ddatblygu, ei ddosbarthu neu ei ddefnyddio. Mae'r dulliau deddfwriaethol yn dibynnu ar fframwaith rheoleiddio sy'n bodoli eisoes, gan gynnwys deddfwriaeth sy'n benodol i un parth neu ddeddfwriaeth sy'n berthnasol i sawl parth.

Er enghraifft, mae'r ddeddfwriaeth bresennol y gellid ei defnyddio i ddeddfu ynghylch camddefnyddio AI cynhyrchiol yng Nghymru yn cynnwys:

- **Deddf Cydraddoldeb 2010**, sy'n amddiffyn pobl rhag gwahaniaethu ac yn cefnogi cyfleoedd cyflogaeth cyfartal;
- **Deddf Diogelu Data 2018**, sy'n gweithredu Rheoliad Cyffredinol yr Undeb Ewropeaidd ar Ddiogelu Data (GDPR) yn y DU; a
- **Deddf Diogelwch Ar-lein 2023**, sy'n cynnwys ei gwneud yn drosedd rhannu cynnwys anweddus ffugio dwfn heb ganiatâd.

Fodd bynnag, cyfyngir ar bwerau deddfu Llywodraeth Cymru a'r Senedd yn y meysydd uchod gan **Ddeddf Llywodraeth Cymru 2006**. Mae diogelu gwybodaeth bersonol yn fater a gedwir yn ôl gan Senedd y DU, ond mae rhywfaint o le i'r Senedd wneud cyfraith sy'n ymwneud â chyfle cyfartal.

Mae **adolygiad barnwrol** yn ffordd o herio cyfreithlondeb penderfyniadau neu weithredoedd a wneir gan gyrff cyhoeddus. Os bydd corff cyhoeddus yn defnyddio AI cynhyrchiol mewn ffordd sy'n gyfystyr â thorri'r gyfraith, gellid defnyddio adolygiad barnwrol i'w ddwyn i gyfrif.

Ffigur 5: I gynnal adolygiad barnwrol, rhaid gwneud cais i'r Uchel Lys Cyfiawnder, sydd wedi'i leoli yn Llundain

Rhwymedigaeth gyfreithiol yw Dyletswydd Cydraddoldeb y Sector Cyhoeddus sy'n ei gwneud yn ofynnol i gyrff cyhoeddus roi sylw dyledus i'r angen i ddileu gwahaniaethu; hyrwyddo cyfle cyfartal; a meithrin perthynas dda rhwng y rhai sydd â nodweddion gwarchodedig a'r rhai nad oes ganddynt nodweddion gwarchodedig. Mae'r **dyletswyddau sy'n benodol i Gymru** yn cynnwys cymryd camau i gyflawni'r ddyletswydd cydraddoldeb gyffredinol uchod ac i gynnal tryloywder; ymgysylltu â grwpiau sydd â nodweddion gwarchodedig; cynnwys gofynion cydraddoldeb mewn ymarferion caffael; cyhoeddi amcanion cydraddoldeb; ac asesu effaith cydraddoldeb. Fodd bynnag **comisiynodd y Comisiwn Cydraddoldeb a Hawliau Dynol waith ymchwil** yn 2022 a ddangosodd nad oedd unrhyw gyrff cyhoeddus yng Nghymru yn cyfeirio at AI na materion digidol yn eu cynlluniau cydraddoldeb strategol cyhoeddedig. Mae'r Comisiwn wedi llunio **canllawiau technegol ar Ddyletswydd Cydraddoldeb y Sector Cyhoeddus i Gymru**. **Mae gan y Comisiwn hefyd bwerau** i gynnal 'asesiad Adran 31', pan fydd yn amau nad yw sefydliad yn cydymffurfio â'r Ddyletswydd, a gall ofyn am dystiolaeth o'r camau y maent wedi'u cymryd i gyflawni'r Ddyletswydd. Os bydd y Comisiwn o'r farn bod y dystiolaeth yn anfoddhaol, gall ofyn i'r sefydliad ymrwymo i 'gytundeb Adran 23' sy'n ei gwneud yn ofynnol **cymryd camau** (e.e. darparu cynllun gweithredu erbyn dyddiad penodol, darparu deunydd dysgu, cyhoeddi

dadansoddiadau, adolygu prosesau, ymgysylltu â rhanddeiliaid penodol). Mae gan y Comisiwn Cydraddoldeb a Hawliau Dynol hefyd bwerau i gynnal ymchwiliadau ac i wneud ymholiadau ynghylch themâu mawr (er enghraifft, gallai'r Comisiwn edrych ar y defnydd o AI mewn cyrff cyhoeddus yng Nghymru) neu sefydliadau penodol.

Daeth y **Ddyletswydd Economaidd-Gymdeithasol** i rym ar 31 Mawrth 2021. Pan fydd cyrff cyhoeddus yn gwneud penderfyniadau strategol ynghylch sut i arfer eu swyddogaethau, mae'r ddyletswydd yn ei gwneud yn ofynnol iddynt roi sylw dyledus i'r angen i leihau anghydraddoldeb canlyniad sy'n deillio o anfantais economaidd-gymdeithasol. Felly, mae angen i gyrff cyhoeddus ystyried a yw'r defnydd o AI cynhyrchiol yn cynyddu neu'n lleihau anghydraddoldebau economaidd-gymdeithasol.

Cymhwysedd ar gyfer deddfwriaeth bellach

Mae'r **Strategaeth Ddigidol i Gymru**, a gyhoeddwyd gan Lywodraeth Cymru ym mis Mawrth 2021, yn tynnu sylw at bwysigrwydd defnyddio AI "mewn ffordd foesegol a chyda gonestrwydd" er mwyn sicrhau "moeseg data, tryloywder a ffydd", a phwysigrwydd hyn i economi Cymru.

Mae diogelu data, yn ogystal ag eiddo deallusol a gwasanaethau rhyngrwyd, yn **faterion a gedwir yn ôl**, sy'n golygu nad oes gan Lywodraeth Cymru a'r Senedd y pŵer i ddeddfu mewn perthynas â'r meysydd hyn a bod Cymru felly'n ddarostyngedig i'r un gyfraith â'r DU. Fodd bynnag, o ran y drefn ehangach o reoliadau diogelu data, **nid yw bob amser yn cynnwys ffin syml rhwng yr hyn 'a gedwir yn ôl' a'r hyn 'sydd wedi'i ddatganoli'**. Mae rhai agweddau ar reoli data wedi'u datganoli (e.e. 'llywodraethu gwybodaeth' ym maes gofal iechyd), mewn rhai meysydd, mae rheoleiddwyr ledled y DU a rheoleiddwyr datganoledig yn rhannu rhywfaint o gyfrifoldeb (e.e. cyfreithiau cydraddoldeb yn yr Alban), ac mewn meysydd eraill, y gwledydd datganoledig sy'n gyfrifol bellach am gyfraith flaenorol yr UE a oedd yn gymwys ledled y DU cyn Brexit. Nododd **Roberts et al** fel a ganlyn:

> (...) each of the devolved nations has different data strategies in place and different rules governing, for example, access to data for secondary purposes including the development of AI.

Gellid defnyddio AI, yn enwedig AI cynhyrchiol sydd â chymwysiadau at ddibenion cyffredinol, mewn amrywiaeth o feysydd a gedwir yn ôl (e.e. diogelu data) a meysydd sydd wedi'u datganoli (e.e. gofal iechyd, addysg), ac oherwydd hynny, **nid yw'n glir** beth fyddai'r drefn o ran cymhwysedd datganoledig ar gyfer deddfwriaeth AI cynhyrchiol yn benodol (ac AI yn gyffredinol). Dywed **Roberts et al** y gallai arwain at densiynau mewnol am fod y gwledydd datganoledig eisiau

cyflwyno deddfwriaeth ynghylch AI, gyda Senedd y DU yn herio hynny ar y sail ei bod yn ymwneud â materion a gedwir yn ôl neu, mewn cyferbyniad, am fod Senedd y DU yn torri Confensiwn Sewel drwy ddeddfu ynghylch AI mewn meysydd sydd o fewn cymhwysedd y Senedd heb gydsyniad y Senedd:

> The ambiguity surrounding how reserved and devolved powers relate to AI creates a real risk of regulatory divergence, which could lead to ineffective protections for citizens on the one hand, and a confusing regulatory environment for companies on the other, undermining a central aim of the UK's overarching approach.

Gwnaethant nodi bod anghysondebau eisoes rhwng y strategaeth a ddewiswyd gan Lywodraeth y DU **sy'n cefnogi arloesedd** a chynllun yr Alban ar gyfer **cenedl ddigidol foesegol** a nodwyd yn ei **Strategaeth AI**, a gyhoeddwyd chwe mis cyn strategaeth y DU.

Nododd **Roberts et al hefyd** nad yw'r Papur Gwyn ar AI a gyhoeddwyd gan Lywodraeth y DU yn mynd i'r afael â'r amwysedd ynghylch pwerau a gedwir yn ôl a phwerau datganoledig ar gyfer AI a'r problemau cydgysylltu posibl a allai ddeillio o hyn (mae rhagor o fanylion isod).

Polisi a deddfwriaeth y DU

Strategaeth ynghylch AI yn benodol

Yn ogystal â'r canllawiau ar gyfer defnyddio AI a nodwyd uchod, mae Llywodraeth y DU wedi cymryd y camau a ganlyn:

- Cyhoeddi **strategaeth AI genedlaethol** ym mis Medi 2021 (diweddarwyd ym mis Rhagfyr 2022), ynghyd â **chynllun gweithredu AI** cysylltieidig a gyhoeddwyd ym mis Gorffennaf 2022;
- Cyhoeddi papur polisi ar **sefydlu dull gweithredu o ran AI sy'n cefnogi arloesedd** ym mis Gorffennaf 2022, sy'n disgrifio'r egwyddorion lefel uchel sydd i'w cymhwyso fesul sector, ac yn nodi nad yw'r camau nesaf yn cynnwys deddfwriaeth newydd. Mae'r papur polisi yn egluro y dylid defnyddio'r pwerau sydd eisoes gan reoleiddwyr i reoleiddio AI, gan ddatganoli pwerau i'r sectorau, gydag egwyddorion traws-sectoraidd anstatudol i ddechrau a chan annog rheoleiddwyr i ddarparu arweiniad a mesurau gwirfoddol yn y lle cyntaf;
- Creu adran newydd, sef yr **Adran Gwyddoniaeth, Arloesedd a Thechnoleg** ym mis **Chwefror 2023**, **sy'n cyfuno** rhannau perthnasol yr Adran Busnes, Ynni a Strategaeth Ddiwydiannol flaenorol a'r cyn Adran dros Dechnoleg Ddigidol, Diwylliant, y Cyfryngau a Chwaraeon, gyda'r nod o hyrwyddo arloesedd a fydd

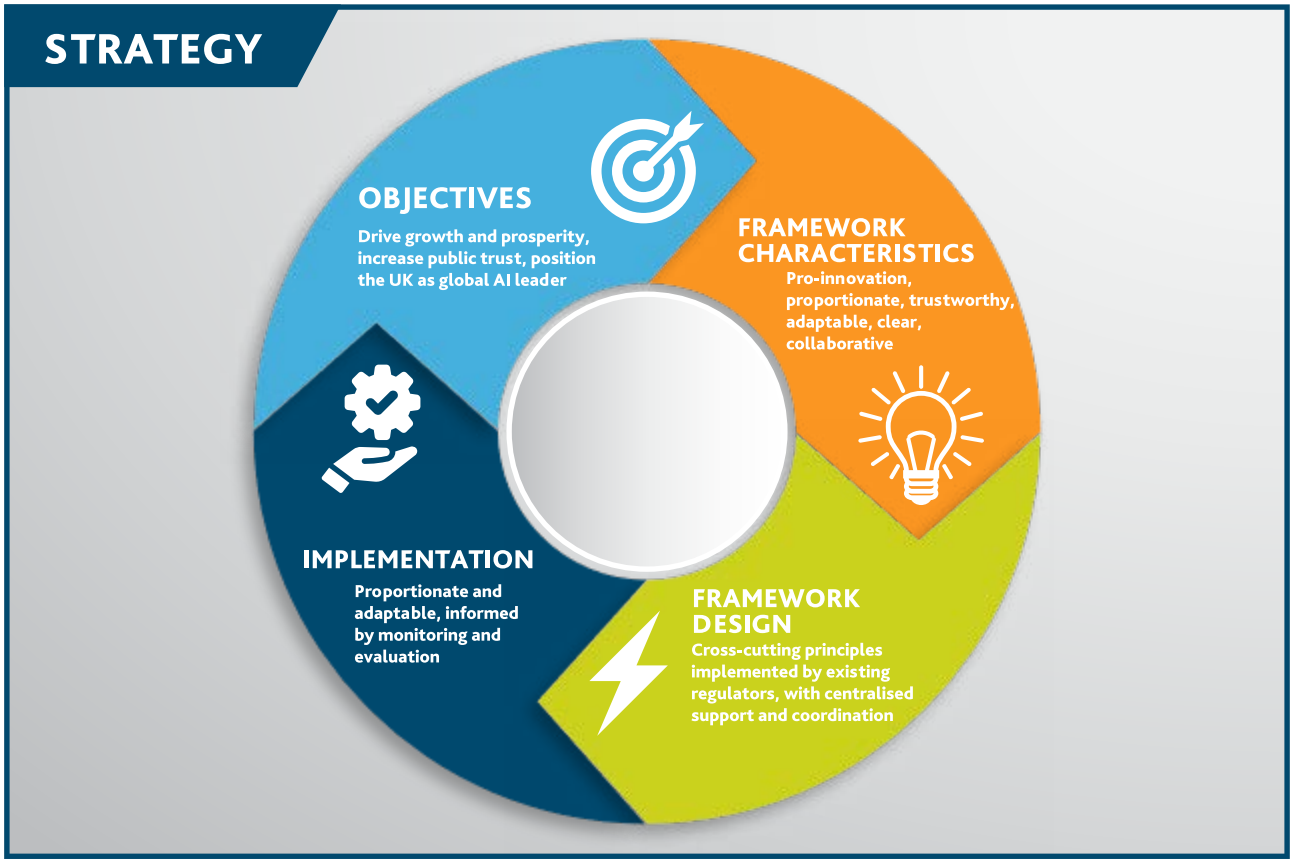
yn darparu gwell gwasanaethau cyhoeddus, creu swyddi newydd sy'n talu'n well ac yn tyfu'r economi;

- Cyhoeddi **Papur Gwyn ar reoleiddio AI** ym mis Mawrth 2023, sy'n seiliedig ar y dull sy'n cefnogi arloesedd a ddisgrifir ym mhapur polisi 2022 a grybwyllir uchod. Mae'r Papur Gwyn yn nodi cynnig Llywodraeth y DU ar gyfer rheoleiddio AI ond nid yw'n cynnwys cynllun i gyflwyno rheoliadau neu gyrff rheoleiddio newydd penodol. Cynhaliwyd yr ymgynghoriad ar gyfer y Papur Gwyn hwn rhwng mis Mawrth a mis Mehefin 2023 a chyhoeddodd Llywodraeth y DU ei **hymateb** ym mis Chwefror 2024. Yn ei hymateb, tynnodd Llywodraeth y DU sylw at y ffaith bod rhanddeiliaid yn cefnogi'r dull, gan bwysleisio bod rheoleiddwyr eisoes yn cymryd camau i ddilyn y fframwaith arfaethedig.
- Darparu **£100 miliwn o gyllid i dasglu modelau sylfaen AI** i arwain gwaith ymchwil hanfodol ar ddiogelwch AI a hybu'r broses o ddatblygu modelau sylfaen mewn modd diogel a dibynadwy, gan hefyd fanteisio ar y cyfleoedd arbennig a gyflwynir yn eu sgil, yn ôl y datganiad i'r wasg o fis Mehefin 2023;
- Arwain **uwchgynhadledd ryngwladol ar ddiogelwch AI** ym mis Tachwedd 2023;
- Cyhoeddi lansiad y **Sefydliad Diogelwch AI** (AISI) ym mis Tachwedd 2023 a darparu **rhagor o wybodaeth** ynghylch ei swyddogaethau a'i ddulliau ym mis Chwefror 2024;
- Cyflwyno'r **Bil Diogelu Data a Gwybodaeth Ddigidol (Rhif 2)**, a gafodd ei ystyried gan Senedd y DU, ac y disgwylid iddo effeithio ar reoleiddio AI gan ei fod yn darparu ar gyfer rheoleiddio prosesu data personol. **Ni fydd y Bil hwn yn symud ymlaen ymhellach** gan fod sesiwn 2023-24 Senedd y DU wedi'i haddoedi;
- Cydnabod, yn ei **ymateb** i'r Papur Gwyn ar reoleiddio AI o fis Chwefror 2024, yr angen am gamau deddfwriaethol ym mhob gwlad unwaith y bydd y ddealltwriaeth o'r risg wedi aeddfedu;
- **Cyhoeddi** buddsoddiad o £80 miliwn i lansio naw canolfan ymchwil newydd ledled y DU i hybu ymchwil AI ym mis Chwefror 2024; a
- Darparu **canllawiau cychwynnol AI** i reoleiddwyr ym mis Chwefror 2024 a gofyn i reoleiddwyr allweddol roi diweddariadau am eu dull strategol o ymdrin ag AI, y cawsant oll eu **cyhoeddi** erbyn diwedd mis Ebrill 2024.

Nododd y Comisiwn Cydraddoldeb a Hawliau Dynol, oherwydd y dull traws-sectoraidd o gefnogi arloesedd a ddewiswyd gan Lywodraeth y DU, y gallai fod yn heriol i reoleiddwyr gael adnoddau ar gyfer y cyfrifoldebau newydd hyn ac i bobl unioni camau am fod cynifer o reoleiddwyr i'w cael.

Nododd yr Ada Lovelace Institute **18 argymhelliad** ar gyfer rheoleiddio AI yn y DU yn dilyn y Papur Gwyn. Cyhoeddwyd yr argymhellion hyn ym mis Gorffennaf 2023. Ym mis Gorffennaf 2023, **cyhoeddodd** Pwyllgor Gwyddoniaeth, Arloesedd a Thechnoleg Tŷ'r Cyffredin ei adroddiad interim ar lywodraethu AI. Yn yr adroddiad hwn, tynnodd y Pwyllgor sylw at risgiau'r dull a gynigiwyd yn y Papur Gwyn a darparodd 22 o argymhellion. Nododd y Pwyllgor y risg o fod ar ei hôl hi wrth i AI a'r fframweithiau o genhedloedd eraill ddatblygu'n gyflym. Gwnaeth hefyd argymell Bil AI ag iddo ffocws penodol iawn i helpu'r DU i sefydlu ei hun fel arweinydd o ran llywodraethu AI. **Ymatebodd** Llywodraeth y DU i'r adroddiad hwn ym mis Tachwedd 2023, gan ailddatgan ei nod o osgoi rhuthro i ddeddfu a nodi y byddai'n rhoi sylw i ddatblygiadau newydd yn y fframwaith rheoleiddio AI yn ei hymateb i'r ymgynghoriad ar y Papur Gwyn.

Ffigur 6: Braslun o strategaeth Llywodraeth y DU ar gyfer rheoleiddio AI



Ffynhonnell: **Llywodraeth y DU**

Goblygiadau'r Papur Gwyn ar reoleiddio AI i Gymru

Nodir yn benodol yn y Papur Gwyn ar AI ei fod yn berthnasol i'r DU gyfan gan fod AI yn effeithio ar ystod eang o sectorau a meysydd polisi, y mae rhai ohonynt wedi'u cadw yn ôl ac eraill wedi'u datganoli. Mae'n mynd ymlaen i esbonio sut y bydd y gweinyddiaethau datganoledig yn rhan o'r gwaith o reoleiddio AI.

Nododd **Roberts et al**, er bod y Papur Gwyn yn hyrwyddo hyblygrwydd yng nghyd-destun yr amgylchedd AI sy'n esblygu'n gyflym, ac er ei fod yn cydnabod pa mor heriol yw cydgysylltu, nid yw serch hynny yn mynd i'r afael â'r tensiynau posibl, na'r heriau o ran capasiti ac ariannu, a allai godi gyda'r llywodraethau datganoledig. Yn ôl Roberts et al, nid yw ychwaith yn darparu unrhyw gamau nesaf neu amserlen gadarn i Lywodraeth y DU gefnogi rheoleiddwyr datganoledig. Awgryma Roberts et al y dylid sefydlu grŵp rhyngweinidogol ar lywodraethu digidol.

Polisi yr UE

Ym mis Ebrill 2021, **cynigiodd y Comisiwn Ewropeaidd** fframwaith cyfreithiol ar AI. Mae'r fframwaith hwn yn seiliedig ar reolau gwahanol yn dibynnu ar lefelau risg: risg annerbyniol, risg uchel, risg gyfyngedig, a risg isel. Amcan y Rheoliad AI, a elwir yn Ddeddf Deallusrwydd Artiffisial (y Ddeddf AI), yw **sicrhau** y bydd y canlynol yn digwydd:

> (...) AI systems used in the EU are safe, transparent, traceable, non-discriminatory and environmentally friendly.

Ym mis Mehefin 2023, **mabwysiadodd Senedd Ewrop ei safbwynt negodi** ar y **Ddeddf AI** gyda 499 pleidlais o blaid, 28 yn erbyn a 93 yn ymatal. Cyhoeddodd **Gwasanaeth Ymchwil Senedd Ewrop bapur briffio** ym mis Mawrth 2024, gan nodi:

European Union lawmakers reached a political agreement on the draft artificial intelligence (AI) act in December 2023. Cafodd y **testun terfynol** ei gymeradwyo ym mis Chwefror 2024, gyda 71 pleidlais o blaid, wyth pleidlais yn erbyn a saith yn ymatal. Ar 21 Mai 2024, **mabwysiadodd y Cyngor Ewropeaidd** Ddeddf AI yr UE yn ffurfiol.

O ran rheoliadau penodol sy'n berthnasol i AI cynhyrchiol, **mae'r Ddeddf AI yn cynnwys**:

- gwahardd arferion AI niweidiol sy'n cael eu hystyried yn fygythiad clir i ddiogelwch, bywoliaethau a hawliau pobl, oherwydd y 'risg annerbyniol' y maent yn ei chreu, sy'n cynnwys systemau AI sy'n defnyddio technegau trin cynnwys isganfyddol niweidiol; systemau AI sy'n manteisio ar grwpiau penodol sy'n agored i niwed (anabledd corfforol neu feddyliol); systemau AI a ddefnyddir gan awdurdodau cyhoeddus, neu ar eu rhan, at ddibenion sgorio cymdeithasol; systemau adnabod biometrig o bell 'amser real' mewn mannau cyhoeddus hygyrch at ddibenion gorfodi'r gyfraith, ac eithrio mewn nifer gyfyngedig o achosion;

- gofynion o ran rheoli risg, profi, cadernid technegol, hyfforddi data a llywodraethu data, tryloywder, goruchwyliaeth ddynol, a seiberddiogelwch ar gyfer systemau AI 'risg uchel' (e.e. systemau a ddefnyddir fel at ddibenion diogelwch ac mewn meysydd penodol fel gorfodi'r gyfraith);
- gofynion tryloywder ar gyfer systemau AI sy'n cyflwyno 'risg gyfyngedig' (e.e. systemau sy'n rhyngweithio â phobl fel sgwrsfotiau a ffugio dwfn);
- dim rhwymedigaethau cyfreithiol ar gyfer systemau AI sydd â 'risg isel neu risg fach iawn'.

Mae'r Ddeddf hefyd yn cynnig rhoi'r dasg o oruchwylio a gweithredu'r rheoliadau i awdurdodau goruchwylio cenedlaethol, rhoi'r dasg o asesu cydymffurfiaeth i awdurdodau gwyliadwriaeth y farchnad genedlaethol, a'r dasg o fonitro rheoliadau ar lefel yr UE i Fwrdd Deallusrwydd Artiffisial Ewropeaidd newydd.

Ffigur 7: Thierry Breton, Comisiynydd Marchnad Fewnol yr UE, Sesiwn Lawn – y Ddeddf AI, Senedd Ewrop yn Strasbwrg, 13 Mehefin 2023

(Ffynhonnell: **Senedd Ewrop**)

Mae papur briffio **Gwasanaeth Ymchwil Senedd Ewrop** yn pwysleisio gwahaniaethau rheoleiddiol a gododd a materion a godwyd ynghylch Deddf AI yr UE, gan gynnwys:

- y ffordd orau o **ddiffinio** systemau AI ac AI at ddibenion cyffredinol;
- **aneffeithiolrwydd** posibl y Ddeddf AI gan fod y categori risg yn dibynnu ar hunanasesiad;
- yr **angen** i gynnwys rhagor o fathau o systemau AI yn y gwaharddiad er mwyn osgoi defnyddio unrhyw system trin cynnwys (e.e. ffugio dwfn) a'r angen am weithdrefn i ychwanegu technolegau at y gwaharddiad;
- yr angen i ddosbarthu risgiau yn fwy **manwl**;
- y **diffyg mecanweithiau** sydd ar gael i ddefnyddwyr ar gyfer gwneud cwynion neu gais i unioni cam drwy brosesau barnwrol;
- **diffyg cysondeb** y Ddeddf AI â'r rheoliadau sectorol presennol, gan osod her i'w gweithredu; a
- cwestiynau ynghylch y rheolau penodol ar gyfer **ceisiadau AI milwrol**.

Gallai'r Ddeddf AI **ddylanwadu** ar bolisi a fabwysiedir gan lywodraethau neu gwmnïau eraill, gan gynnwys yn y DU, drwy 'effaith Brwsel'. Mae hyn yn disgrifio ffenomen lle mae gallu a maint marchnad yr UE yn dylanwadu ar gwmnïau i ddilyn ei safonau ac yn ei dro, mae cwmnïau'n pwyso ar lywodraethau eraill i gydymffurfio â rheolau'r UE. Dyma a ddigwyddodd gyda'r Rheoliad Cyffredinol ar Ddiogelu Data (GDPR).

Ym mis Medi 2022, cynigiodd Comisiwn yr UE hefyd **Gyfarwyddeb Atebolrwydd Deallusrwydd Artiffisial (y Gyfarwyddeb Atebolrwydd AI)** ac **adolygiad o'r Gyfarwyddeb Atebolrwydd am Gynhyrchion**. Nododd **beirniaid** fodd bynnag, fod angen gwelliannau i sicrhau trefniadau rheoleiddio cyson, i gynnwys ystyriaethau cynaliadwyedd, ac i ddiffinio atebolrwydd AI yn well.

Gwledydd eraill

Mae **mwy nag 800 o fentrau polisi AI** ledled y byd.

Ers mis Mai 2023, UDA sydd â'r nifer uchaf o bolisïau AI ar lefel genedlaethol. Nod **Deddf Menter Deallusrwydd Artiffisial Cenedlaethol 2020** UDA oedd cyflymu gwaith ymchwil a chymhwyso o ran AI, ac o dan y Ddeddf hon sefydlwyd Swyddfa Menter AI Cenedlaethol, Pwyllgor Cynghori AI Cenedlaethol a Thasglu Adnoddau Ymchwil AI Cenedlaethol. Cyhoeddodd y Tŷ Gwyn **lasbrint ar gyfer Bil Hawliau AI** ym mis Hydref 2022, sy'n cynnwys canllawiau nad ydynt yn rhwymol ar ddefnyddio AI yn ddiogel ac yn effeithiol, gyda mesurau i ddiogelu rhag gwahaniaethu algorithmig ac ystyriaethau preifatrwydd data.

Rhyddhaodd y National Institute of Standards and Technology yn UDA **Fframwaith**

Rheoli Risg AI ym mis Ionawr 2023 a lansiodd y **Trustworthy and Responsible AI Resource Center** ym mis Mawrth 2023. Ym mis Gorffennaf 2023, **sicrhaodd y Tŷ Gwyn ymrwymiad gwirfoddol** gan gwmnïau AI blaenllaw o ran rheoli risgiau AI.

Ym mis Hydref 2023, llofnododd Arlywydd UDA **Orchymyn Gweithredol ar ddatblygu a defnyddio AI yn ddiogel ac yn ddibynadwy**. Ym mis Ebrill 2024, **cyhoeddodd** asiantaethau ffederal eu bod wedi cwblhau'r camau gweithredu yn y Gorchymyn Gweithredol.

Mae UDA hefyd wedi bod yn cydweithio â'r UE, er enghraifft drwy **Gyngor Masnach a Thechnoleg yr UE ac UDA** i lunio **map trywydd AI ar y cyd**. Cytunodd y DU ac UDA ar **femorandwm cyd-ddealltwriaeth** diogelwch AI ym mis Ebrill 2024. Soniodd **Roberts et al** am botensial 'effaith drawsatlantig' yn hytrach nag 'effaith Brwsel' ar y DU.

Fodd bynnag, nid yw UDA wedi cyflwyno unrhyw drefn reoleiddio genedlaethol. Mae gwladwriaethau ac awdurdodau dinasoedd **yn datblygu** eu rheoliadau eu hunain. Er enghraifft, cyhoeddodd Dinas Efrog Newydd **gynllun gweithredu ar AI** ym mis Hydref 2023.

Mae gwledydd eraill wedi bod yn datblygu rheoliadau AI:

- Mae Brasil wedi cyhoeddi **strategaeth genedlaethol ar AI** ac wedi **cyflwyno rheoliad AI drafft** sy'n **hyrwyddo** dull sy'n debyg i Ddeddf yr UE. Mae'r rheoliad hwn yn **ei gwneud yn ofynnol** i ddarparwyr AI ddarparu gwybodaeth i ddefnyddwyr, rhoi pŵer i ddefnyddwyr ofyn am ragor o wybodaeth neu herio penderfyniad AI, ei gwneud yn ofynnol i ddatblygwyr gynnal asesiadau risg, ac yn gwneud datblygwyr yn atebol am niwed a achosir gan eu systemau;
- Mae Tsieina wedi cyhoeddi **mesurau interim ar gyfer rheoli gwasanaethau AI cynhyrchiol** a ddaeth i rym ym mis Awst 2023. Mae'r rhain yn nodi mai cynhyrchwyr y cynnwys sy'n gyfrifol am y cynnwys a gynhyrchir. Maent hefyd yn nodi bod angen i ddarparwyr sicrhau bod y ffynonellau a ddefnyddir i hyfforddi AI cynhyrchiol yn gyfreithlon, a bod cyfyngiadau ar natur y ffynonellau hyn, a bod rhaid iddynt beidio â thorri hawliau eiddo deallusol. Mae'r mesurau hyn hefyd yn ei gwneud yn ofynnol bod asesiadau diogelwch yn cael eu cynnal cyn defnyddio cynhyrchion AI cynhyrchiol i ddarparu gwasanaethau i'r cyhoedd. Yn olaf, mae'n dweud y dylai'r darparwyr ddefnyddio mesurau effeithiol i wella tryloywder gwasanaethau AI cynhyrchiol ac i wella cywirdeb a dibynadwyedd y cynnwys a gynhyrchir. Mae Tsieina hefyd yn gweithio ar **gyfraith AI cyffredinol** y disgwylir y bydd drafft ohoni'n barod yn 2024; ac

- Mae Canada wedi cynnig **Deddf Deallusrwydd Artiffisial a Data** sy'n **dilyn** trywydd tebyg i'r UE. Byddai'r Ddeddf yn ei gwneud yn ofynnol i ddatblygwyr gael llifoedd gwaith i leihau risgiau, gwella tryloywder, sicrhau parch at gyfreithiau gwrth-wahaniaethu, yn ogystal â phrosesau gwneud penderfyniadau clir. Ers mis Mai 2024, mae'r Bil **dan ystyriaeth** yn Nhŷ'r Cyffredin.

Mae rhai gwledydd sydd wedi llunio strategaethau ac wedi cyhoeddi canllawiau, ond nad oes ganddynt gynlluniau ar gyfer rheoliadau. Er enghraifft:

- Mae'r Emiraethau Arabaidd Unedig wedi cyhoeddi **strategaeth genedlaethol ar gyfer deallusrwydd artiffisial**, ond nid yw'r strategaeth hon yn cynnwys cynlluniau ar gyfer rheoliadau ar wahân i adolygu dulliau cenedlaethol o ymdrin â materion fel rheoli data, moeseg, a seiberddiogelwch a'r arferion gorau rhyngwladol diweddaraf o ran deddfu ynghylch risgiau byd-eang yn sgil AI;
- Nid oes gan Japan **unrhyw reoliadau** sy'n benodol i AI a defnyddir cyfreithiau cysylltiedig (e.e. ar gyfer diogelu data) yn lle hynny. Mae gan Japan **Ddeddf Hawlfraint** ddiwygiedig sy'n awdurdodi'r defnydd o ddata hawlfraint ar gyfer dadansoddi data, gan gynnwys eu defnyddio i hyfforddi AI. Fodd bynnag, ym mis Mai 2024, **cyhoeddodd** Prif Weinidog Japan fframwaith rhyngwladol gwirfoddol i reoleiddio AI cynhyrchiol o'r enw yr Hiroshima AI Process Friends Group;
- Mae India **wedi datblygu** strategaeth genedlaethol ar AI, cyhoeddi canllawiau ar AI cyfrifol, lansio tasglu ar gyfer trawsnewid economi India, lansio rhaglen genedlaethol ar AI (India AI), ac yn ddiweddar **cyhoeddodd** ofyniad i gwmnïau technoleg 'o bwys' gael cymeradwyaeth y llywodraeth cyn rhyddhau modelau newydd, ond nid yw hyn yn gyfreithiol rwymol ac fel y mae ym mis Mai 2024, **nid yw India wedi** cyflwyno rheoliadau; ac
- Mae Awstralia **wedi hyrwyddo** wyth egwyddor moeseg AI a fframwaith gwirfoddol AI i sicrhau bod AI yn ddiogel ac yn ddibynadwy, ond **dim rheoliadau**.

Mae'r pedair gwlad hyn wedi llofnodi **datganiad Bletchley** ym mis Tachwedd 2023 ac felly maent yn cydnabod agenda i weithredu polisïau sy'n seiliedig ar risg (gweler rhagor o fanylion isod).

4. Beth yw'r heriau o ran cynllunio a gweithredu polisi?

Esblygu'n gyflym

Gan fod y sefyllfa o ran AI cynhyrchiol yn esblygu'n gyflym, mae **angen i reoliadau fod yn hyblyg** i gyfrif am ddatblygiadau newydd. Mae academyddion a rheoleiddwyr felly'n **awgrymu** y dylid llunio deddfau sy'n niwtral o safbwynt technoleg, gan ganolbwyntio yn lle hynny ar gymwysiadau penodol sy'n peri risg uchel.

Yn ogystal â'r angen i fod yn hyblyg, mae **angen i reoliadau hefyd** ddarparu digon o eglurder cyfreithiol iddynt fod yn effeithiol. At hynny, mae **angen** iddynt fod yn ddigon cwmpasog iddynt osgoi bylchau cyfreithiol ac mae angen iddynt gael eu mabwysiadu'n gyflym.

Felly, mae **angen** sicrhau bod rheoleiddwyr, cymdeithas sifil a sefydliadau yn cael pwerau ac adnoddau priodol i allu addasu i'r amgylchedd AI hwn sy'n esblygu'n gyflym.

Diffinio niwed ac asesu difrod

Nid yw bob amser yn hawdd diffinio effeithiau AI cynhyrchiol, fel y nodwyd yn yr Harvard Business Review. Gall fod yn anodd diffinio pwy sydd wedi cael ei niweidio a phryd, er enghraifft gydag allbynnau anghywir. At hynny, gallai ymddangos mai dim ond ôl-effaith fach sydd i gamgymeriad bach, ond wrth i'r effaith ledaenu gall arwain at ganlyniadau cymdeithasol niweidiol. Mae'n arwain at drafferthion wrth nodi'r hyn sy'n 'ddigon niweidiol' i'w ystyried yn anghyfreithlon.

O ganlyniad, gall fod yn heriol asesu'r canlyniadau a'r gosb briodol hefyd. Gall fod yn anodd pennu'r cosbau i'w rhoi a phryd y dylid eu rhoi. Gwelwyd problemau tebyg gyda materion preifatrwydd.

At hynny, os ystyrir bod AI cynhyrchiol yn 'lleferydd', gall arwain at heriau ychwanegol sy'n gysylltiedig ag anawsterau wrth reoleiddio lleferydd. **Nodir** fel a ganlyn mewn erthygl gan Brifysgol Manceinion:

> Regulating online content can be problematic due to challenges, such as defining the legally responsible actors for online hate speech and balancing a right to free speech with the desire to limit harmful content.

Yr angen am ddulliau rhyngwladol

Cydnabyddir bod angen cydweithredu rhyngwladol er mwyn osgoi trosglwyddo'r problemau i fannau eraill. Mae'r cyfryngau cymdeithasol yn enghraifft lle nad oes strategaeth gyffredin wedi'i gweithredu i wahardd sylwadau difrïol a gwahaniaethol.

Llofnodwyd **Datganiad Bletchley** gan 28 o wledydd a'r Undeb Ewropeaidd ym mis Tachwedd 2023. Mae'r datganiad hwn yn cydnabod bod llawer o risgiau sy'n rhyngwladol eu natur wrth raid, ac felly mai'r ffordd orau o fynd i'r afael â hwy yw drwy gydweithrediad rhyngwladol. Ymrwymodd y llofnodwyr i gydweithredu ar egwyddorion cyffredin a chodau ymddygiad, er enghraifft drwy uwchgynadleddau diogelwch AI rhyngwladol yn y dyfodol ac adeiladu dealltwriaeth wyddonol ar sail tystiolaeth a rennir o'r risgiau sy'n gysylltiedig ag AI. Fodd bynnag, maent hefyd yn cydnabod efallai bod eu dulliau yn wahanol oherwydd eu hamgylchiadau cenedlaethol a'r fframweithiau cyfreithiol sy'n gymwys. Cynhaliwyd yr **uwchgynhadledd diogelwch AI ddiweddaraf** ym mis Mai 2024 yn Seoul, yng Ngweriniaeth Korea.

Ym mis Mawrth 2024, llofnododd Gweinidogion technoleg a digidol y G7 **ddatganiad** ar ddatblygu pecyn cymorth AI ar gyfer y sector cyhoeddus.

Mae panel cynghori arbenigol rhyngwladol **wedi cael y dasg** o gynhyrchu adroddiad gwyddonol rhyngwladol ar ddiogelwch AI, er mwyn trafod ble a pham y mae anghytundeb yn y gymuned arbenigol ehangach, ac i gyflwyno'r ddadl mewn modd gwrthrychol i'r gymuned arbenigol. Disgwylir yr adroddiad llawn tua diwedd 2024, gyda fersiwn gychwynnol yn cael ei chyhoeddi yn ystod Gwanwyn 2024.

Termau

Cymraeg	**Saesneg**
Deallusrwydd Artiffisial (AI)	Artificial Intelligence (AI)
Y Swyddfa Digidol a Data Ganolog	Central Digital and Data Office (CDDO)
Y Ganolfan Gwasanaethau Cyhoeddus Digidol	Centre for Digital Public Services (CDPS)
Dysgu Dwfn	Deep Learning (DL)
Asesiadau o'r Effaith ar Ddiogelu Data	Data Protection Impact Assessment (DPIA)
Yr Adran Gwyddoniaeth, Arloesedd a Thechnoleg	Department for Science, Innovation and Technology (DSIT)
Deallusrwydd Artiffisial Cynhyrchiol (AI Cynhyrchiol)	Generative Artificial Intelligence (GenAI)
Model Iaith Mawr (LLM)	Large Language Model (LLM)
Dysgu Peirianyddol	Machine Learning (ML)
Y Ganolfan Genedlaethol Seiberddiogelwch	National Cyber Security Centre (NCSC)
Prosesu Iaith Naturiol	Natural Language Processing (NLP)
Rhwydwaith Niwral	Neural Network (NN)
Dyletswydd Cydraddoldeb y Sector Cyhoeddus	Public Sector Equality Duty (PSED)
Atgyfnerthu'r hyn a ddysgir drwy Adborth Dynol	Reinforcement Learning from Human Feedback (RLHF)
Y ddyletswydd economaidd-gymdeithasol	Socio-economic Duty (SED)