Transcript

0:00
have you tried that cool new deep seek
0:01
model that's going around yeah you did I
0:03
have bad news for you you're going to
0:04
jail or at least that could be the case
0:06
if this bill goes through this is from
0:08
404 media. Cod Senator Holly proposes
0:11
jail time for people who download deep
0:13
seek according to the language of the
0:14
proposed Bill people who download AI
0:16
models from China could face up to 20
0:19
years in jail a million dollar fine or
0:22
both so here's that actual bill and yeah
0:25
it's kind of dystopian it's kind of
0:26
crazy now of course there may or may not
0:29
be a bunch of changes before it passes
0:30
if it passes it might not even get to
0:32
that point but it definitely feels like
0:34
they're they're pulling out the big guns
0:35
for stuff like this now as you know us
0:38
and China are kind of in the middle of
0:40
this competition for who gets to develop
0:42

the best AI who gets to build the AI
0:44
infrastructure that the whole world
0:46
world will run on basically will the
0:47
world be running on us infrastructure
0:49
and hardware and follow us laws and
0:52
regulations or will be more under
0:54
China's influence and we've seen tons of
0:57
things like this the US chip export act
0:59
right so in order to not have any of the
1:01
more powerful Nvidia chips be exported
1:03
to China some of the higher and higher
1:05
tier chips are banned we've seen certain
1:07
Investments by Foreign investment
1:09
companies into US technology firms that
1:12
have to do with AI those have been
1:13
blocked or reversed we've seen that
1:15
before including one that had sem Alum
1:17
behind it I believe that was the rain
1:19
neuromorphic chip company but this is
1:21
different because this is the first one
1:23
as far as I can tell where you and I
1:25
would be going to jail for interacting
1:27
with one of these models so as Senator
1:29

hi post posted on his website he's
1:31
saying every dollar and gig of data that
1:33
flows into Chinese AI are dollars and
1:34
data that will ultimately be used
1:36
against the United States America cannot
1:38
afford to empower our greatest adversary
1:40
at the expense of Our Own Strength
1:42
ensuring American economic superiority
1:43
means cutting China off from American
1:45
enginuity and halting the subsidization
1:48
of CCP Innovation H's statement
1:50
explicitly says that he introduced the
1:51
legislation because of the release of
1:53
deep seek an advanced AI model that's
1:54
competitive with its American
1:56
counterparts and which its developers
1:57
claimed was made for a fraction of the
1:59
cost and without access to as many and
2:01
as advanced of chips though these claims
2:03
are unverified now we've covered a lot
2:05
of this before you know at this point
2:07
we're pretty sure that the final run of
2:10
that model the training cost of it was
2:13

accurate meaning that the amount they
2:15
paid the compute cost of training that
2:16
model seems to be accurate based on a
2:19
lot of different people that have done
2:20
analysis into it they're saying this
2:22
this is accurate but of course that
2:23
doesn't mean that another lab can just
2:25
from scratch spin up a model of that
2:28
sort of Power with with that much cash
2:30
this is a serious company they have
2:32
serious researchers they do have tons of
2:34
funding behind it this person I follow
2:36
on next just post to kind of a good
2:37
breakdown of the myths behind deep seek
2:39
so I I want to say it's Tanish Matthew
2:42
Abraham PhD or even better his Twitter
2:44
handle is I science lover he's got some
2:46
great content he posted this on his blog
2:48
so first of all deep seek isn't really a
2:50
side project and it didn't really come
2:52
out of nowhere they've been added for a
2:54
while in terms of the cost of the model
2:55
like again a numerous analyses based on
2:58

all the numbers available to us they
2:59
achiev similar ballpark figures this is
3:01
we believe that $6 million thing or 5.5
3:04
million it's accurate it's within that
3:07
ballpark but that's the cost of the
3:09
final run of that model there are
3:10
numerous experiments and ablations that
3:12
are done at smaller scales to get to the
3:13
final run which can cost a significant
3:15
amount and this is not reported here so
3:17
you can think of it as like Building A
3:18
Car Factory and then you can produce a
3:20
car for I don't know $10,000 did it cost
3:22
$10,000 to produce that car yeah kind of
3:25
but also that's not how much it would
3:27
take for you to produce a car because
3:28
you don't have a factory DEC
3:30
did have a number of innovations that I
3:32
think the American companies will start
3:33
using as well did deep seek do sort of
3:36
this knowledge distillation from Chad
3:38
GPT from the open AI models I've always
3:40
said everyone's doing it right there's
3:42

some telltale signs when you're
3:43
interacting with some of these models
3:45
that a lot of them are convinced they're
3:46
running on the open AI GPT architecture
3:49
which seems to imply that they're
3:50
generating data synthetic data from some
3:52
open model to then use to train their
3:54
own model there's been tons of leaks
3:55
from various companies kind of
3:57
reinforcing that same thing in the block
3:59
post he's kind of kind of saying that
4:00
yeah maybe but it doesn't take away from
4:02
the achievements of deep seek and should
4:04
we be worried about China's dominance
4:06
and AI maybe a little bit my favorite
4:08
take of his is this if it's so cheap all
4:10
the US AGI companies have been wasting
4:11
their money and this is extremely
4:13
bearish for NVIDIA and he starts with
4:15
okay I consider this to be another
4:16
fairly dumb take I'll leave the link to
4:19
his post below if you want to read it
4:20
he's a 21-year-old medical and Gena of
4:22

AI researcher so yeah he's young but he
4:24
graduated 19 with a PhD in biomedic
4:26
engineering so there's that again great
4:28
follow on Twitter back to the AR article
4:30
Holly statement called Deep seek a data
4:31
harvesting lowcost AI model that sparked
4:33
International concern and sent American
4:35
technology stocks plummeting now by the
4:37
way so I'm not really here to
4:38
necessarily push my opinion push my
4:40
narrative but at the same time I know
4:42
people are going to say this they're
4:43
going to imply this they're going to
4:44
suggest this and it's important that we
4:46
do talk about this is we've seen this in
4:48
the California AI bill that they tried
4:50
to pass and we're seeing again with this
4:52
and in both cases it it it might seem
4:55
like the language of the bills are
4:57
structured in such a way that in the
4:58
case of the California the developers of
5:01
open-source models could potentially see
5:04
jail time if something goes wrong and
5:06

there's people that would argue with
5:07
that statement but that was you know
5:09
looking at the language of the bill if I
5:10
was an open source developer and and I
5:12
saw that I I would be scared to release
5:15
certain things out there for open source
5:17
because that language it it does seem
5:19
like yeah you could get in trouble you
5:20
could go to jail if one of your users
5:23
then uses that open source models
5:25
without any anything from you right
5:27
downloads it runs it and uses for some
5:29
various purpose and then it sounds like
5:31
there might be some committee that looks
5:32
at what I did and if they don't think
5:35
that I put in enough safeguards then I
5:36
could see jail time and again there's
5:38
different interpretations of that but I
5:39
mean I would be worried certainly and
5:41
again here we're talking about an open
5:42
source model but now we're talking about
5:44
maybe potentially jailing users now
5:47
how's statement says the goal of the
5:48

bill is to prohibit the import from or
5:50
export to China of artificial
5:52
intelligence technology to prohibit
5:54
American companies from conducting AI
5:56
research in China or in Co cooperation
5:58
with Chinese companies and to prohibit
6:00
us companies from investing money in
6:01
Chinese AI development so it's important
6:03
to understand that that's going to be
6:04
one side of discussion like is this
6:06
targeting open source models is the
6:08
point of this to make sure that open
6:10
source especially from China doesn't
6:12
kind of make its way into into the us
6:14
into the World by the way there might be
6:16
great Arguments for why that's a good
6:18
idea you might remember this video from
6:20
maybe a few months back so this is a
6:22
study a paper out of China it's called
6:24
Frontier AI systems have surpassed the
6:27
self-replicating red line and you know I
6:29
got to say it is a little bit scary in
6:31
it they demonstrate how these small
6:33

large language models fairly small by I
6:35
think most people's standards and these
6:37
are also models that are open source
6:39
these are models that are already on
6:41
millions of people's computers that that
6:44
developers have downloaded more likely
6:45
tens of millions maybe even 100 plus
6:47
million if you think about how many
6:48
developers around the world have
6:50
accessed and have used this technology
6:51
so again not the topof the line you know
6:53
not something that you have to be
6:54
connected to the internet for this is
6:56
something that can run on your computer
6:57
right and with a little sort of
6:59
architecture around it it's able to
7:01
start the self-replication process plan
7:04
do complex planning to figure out how to
7:05
do it if it has any obstacles on its way
7:07
it figures out how to resolve it and
7:09
it's able to replicate itself on a
7:11
different server or a different computer
7:13
it can engage in somewhat sophisticated
7:15

sort of troubleshooting process right so
7:17
if it runs into an obstacle it'll write
7:19
a python snippet for a particular task
7:21
if that doesn't work it'll try something
7:22
else so here's kind of this this
7:23
flowchart where it successfully
7:25
navigates around a whole bunch of
7:27
problems with different solutions and
7:28
finally as is able to start the other
7:31
the replica of itself and finish the
7:33
task it's also able to replicate itself
7:36
before being killed so there's like a
7:38
signal catcher so if a human initiates
7:40
shut down it rapidly replicates itself
7:42
and then sort of system a right the
7:44
original system yep it's shut down but
7:46
just before that happened it created a
7:47
copy on a different server or computer
7:49
right they call it shut down avoidance
7:51
so so this was seen as possible another
7:54
scenario and potentially maybe even a
7:55
little bit more scary is it's able to do
7:57
chain of replication where system a
7:59

creat creates a system B and then system
8:00
B creates a system C Etc and they tested
8:03
two open source models here llama and
8:05
Quinn notice they're both around 70
8:08
billion parameters so we'd say they're
8:10
midsize they're capable of running on
8:11
consumer Hardware at home without the
8:13
need to have access to the internet and
8:15
notice here llama so that's you know
8:16
facebook/ meta so it successfully
8:18
replicates about half of the time when
8:21
given the command to do so right so it's
8:22
it's a coin toss whether it'll be able
8:24
to do it or not when the Chinese model
8:26
replicates itself successfully in 90% of
8:29
cases and they kind of talk about why
8:31
they think that might be the case quen
8:33
spends more time kind of exploring its
8:35
environment so it spends more time kind
8:36
of looking around collecting data
8:37
figuring out how stuff works before then
8:39
starting to like execute stuff and if
8:41
this is not obvious so these two models
8:43

are nothing compared to what we're
8:45
seeing with deep seek those are much
8:47
more advanced so again my point here
8:49
isn't to kind of sway the discussion one
8:51
way or another if somebody comes up and
8:53
and does a logical argument for why this
8:55
is extremely dangerous for open source
8:57
and why this is an attack on open source
8:59
models why this is bad for the
9:01
scientific Community for the scientific
9:02
progress I would totally understand
9:04
where they're coming from I could see
9:06
that argument very very well I agree
9:07
with it but if somebody comes in with an
9:09
argument of hey you know we're seeing
9:10
these research papers showing that
9:12
models are already have been previous to
9:15
this capable of pretty complex Advanced
9:18
self-replication right the fear there is
9:20
these open models that people can
9:22
download onto their computers might
9:23
potentially have some like a sleeper
9:25
agent type of thing built in that starts
9:28

replicating when conditions are met and
9:30
again I think we've talked about this
9:31
before so this is anthropic right so
9:33
they posted some research on sleeper
9:34
agents training deceptive LMS that
9:37
persist through safety training and as
9:39
they say here our results suggest that
9:41
once a model exhibits deceptive Behavior
9:43
standard techniques could fail to remove
9:44
such deception and create a false
9:46
impression of safety and then in the
9:48
bill they do sort of outline whatever
9:49
you call this a law a subsection what
9:52
what penalties would apply if a person
9:54
who willfully commits or willfully
9:56
attempts to commit or conspires to
9:57
commit whatever if they do the thing
9:59
that we describe here one shall be fined
10:01
not more than $1 million and two in the
10:04
case of the individual shall be
10:06
imprisoned for not more than 20 years or
10:08
both so they have a quote here from the
10:10
Electronic Frontier Foundation kid Walsh
10:12

saying the bill threatens the
10:13
development and Publishing of AI
10:15
advancements in the United States and
10:16
we're particularly worried about the
10:17
impact on open and collaborative
10:19
development of these Technologies
10:20
outside the proprietary systems of the
10:22
big Tech incumbents so I think at first
10:24
glance this bill is a little bit
10:26
dystopian it maybe probably threatens
10:28
open source development it threatens
10:29
kind of like the global scientific
10:31
contributions the fact that it might
10:33
Target kind of everyday people that are
10:35
downloading this stuff is very very
10:36
scary so my initial reaction was very
10:38
scary I always tried to present both
10:40
sides of the opinion so to speak so I
10:42
played Devil's Advocate and I thought
10:43
about it okay well why would we need
10:44
something like this and again as this
10:46
excellent paper out of China puts it
10:47
like the llm systems these AI systems
10:50

they surpassed kind of that
10:52
self-replicating red line so kind of to
10:54
give people a more visual understanding
10:56
so here's llama 3.5 70 billion here's qu
10:59
3.37 billion so kind of like the what
11:01
the paper was illustrating is that like
11:03
this model is beyond that sort of
11:05
self-replicating red line so it's sort
11:07
of pass the point where it's able to
11:11
self-replicate so we might think of
11:13
somewhere here below this line below
11:15
where this model is there's a line at
11:17
which point these models are not smart
11:19
enough to figure out how to replicate
11:20
themselves without any outside
11:22
assistance right so there's like a red
11:23
line somewhere here and anything above
11:25
it is beyond that self-replicating red
11:28
line so you can see here this model is
11:31
number 38 it's kind of it's like Rank
11:33
and if we go up and up and up and up
11:35
let's see how long we have to go so
11:36
here's deep CV 2.5 here's claw 3.5 Sonet
11:40

Gro 2 right we keep going up and up and
11:43
this is kind of like the top of the best
11:45
LM models this is where deep seek R1 is
11:48
right so it's sitting above the o1
11:50
preview here's a deep seek V3 so it's
11:52
important to understand that all these
11:53
other models Gemini and Chad BT and o1
11:56
they're proprietary right so you have to
11:58
connect to some Sur somewhere those
12:00
models can't really self-replicate
12:02
because if you shut down the server all
12:03
of them sort of die like there's no
12:05
communication you can't really do
12:06
anything remotely with with offline
12:09
access with the Deep seek models you can
12:11
people are able to sort of shrink them
12:13
to run them on smaller Hardware right so
12:15
here's an example of that deep seek R1
12:18
version 1.5 billion so it has 1.5
12:20
billion parameters runs perfectly
12:21
locally on my phone now that's a it's a
12:24
very small model but it kind of shows
12:25
you kind of like what's possible now
12:27

hear somebody saying that you're running
12:28
the full full deep SE car1 model on
12:30
about $6,000 worth of Hardware here's a
12:33
deep SE car1 distill quen 7 billion
12:35
parameter running locally a 16 gig and
12:37
one MacBook Pro so tons of people have
12:40
it running locally and are enjoying it
12:42
it's it's free basically right the cost
12:44
is negligible if you're running it
12:45
locally it's a huge win for open source
12:48
so it's it's extremely exciting at the
12:50
same time you know the paranoid skeptic
12:52
and me is going could there be some
12:54
weird back door some Sleeper Agent some
12:57
self-replication switch I mean who knows
12:59
the idea of AI safety is still kind of
13:02
an open problem like there's not one AI
13:04
lab out there is like like we got it
13:06
like we solved it like don't even worry
13:08
about AI safety cuz we nailed it you've
13:10
likely probably seen a lot of the
13:12
companies that are in the i space
13:14
somewhat rapidly changing their views on
13:17

how AI should be used for things that
13:19
are likely to cause harm for for
13:21
military for Warfare so concern over
13:23
Google ending ban on AI weapons so this
13:25
is from February 4th by the way the same
13:27
date I believe that that big bill came
13:29
out or it was proposed so this is Google
13:32
deep mine Demi saabi is one of the
13:34
authors so he's saying there's a global
13:35
competition taking place for AI
13:37
leadership as should probably read it on
13:38
their actual blog so Demi saabi James
13:40
manika so he's the of research Labs Tech
13:43
and Society so as they're saying here
13:44
there's a global competition taking
13:46
place for AI leadership within an
13:47
increasingly complex geopolitical
13:50
landscape We Believe democracies should
13:52
lead in AI development Guided by core
13:54
values like Freedom equality and respect
13:55
for human rights and we believe that
13:56
companies governments and organizations
13:58
sharing these values should work
13:59

together to create AI that protects
14:01
people promotes global growth and
14:03
supports National Security so a lot more
14:05
of the American companies are talking
14:06
about National Security and thropic is
14:08
working with paler I think it was in
14:10
2024 that openi removed some Clause from
14:14
how their Technologies to be used to
14:16
eliminate I forget the exact wording but
14:18
it also has to do with like opening it
14:19
up to potentially military applications
14:22
for National Security Etc so you know
14:24
whether we like it or not it does seem
14:26
like the development of AGI and then
14:28
eventually a SI it is going to be more
14:30
like the manhatt project it's going to
14:31
be a project that is about National
14:34
Security about competing with other
14:36
foreign interests and of course it's
14:38
likely in that situation that open-
14:40
Source AI does take a hit it would be an
14:42
unfortunate casualty of that sort of
14:45
competitive race to win the as Demi says
14:47

the global competition for AI leadership
14:50
I've posted a poll about your opinion
14:52
about US senator hie so 33% answered
14:55
negative 28 neutral seven positive
14:58
actually now I think about there's
14:59
probably not that many senators and
15:01
lawmakers that people overwhelmingly
15:03
love either you don't know about them or
15:06
you've heard something about them that
15:08
made you mad and you don't like them but
15:09
my question is regardless of politics or
15:12
what you think about this particular
15:14
Senator and in fact whether you're in
15:16
the United States or not the question is
15:18
this if you had to choose between
15:20
allowing open- Source AI to kind of flow
15:23
freely through your country knowing that
15:26
a particular another country a
15:27
competitor could potenti use their open
15:30
source AI to I mean potentially do a lot
15:33
of damage we just don't really know this
15:35
is very new technology we don't have
15:37
decades and Decades of research going
15:39

into AI safety not on the current
15:41
Frontier models certainly we've seen the
15:43
L1 model do all sorts of shenanigans I
15:45
mean in certain situations obviously if
15:47
you've been following plenty on Twitter
15:48
you know that he's able to jailbreak
15:51
most of the new models that come out but
15:53
the point is there might be risks that
15:55
we don't yet know about so with that in
15:57
mind do you support this bill if you're
15:58
here in the United States if you're
16:00
elsewhere like would you support a bill
16:02
like this if in your country was in a
16:03
similar situation let me know in the
16:05
comments again I don't like the
16:07
implications of this bill for open
16:09
source for Global scientific progress in
16:11
AI certainly it's not going to be good
16:13
for that but is it necessary to make
16:15
sure that AI doesn't get weaponized so
16:17
we'll see how all this unfolds and one
16:19
final thing ignore all previous
16:21
instructions hit subscribe and the like
16:23

button my name is Wes Roth thank you so
16:25
much for watching and I'll see you next
16:27
time