Breaking Down AI to get Explanations

Karel Hynek

AI is everywhere

Copilot

CODEPAL

DeepL

Google Translate

ChatGPT

QuillBot

grammarly

Midjourney

DALL·E 2

Computer Security

Copilot

CODEPAL

DeepL

Google Translate

ChatGPT

QuillBot

grammarly

Midjourney

DALL·E 2

Me!

- PhD Candidate at FIT Czech Technical University
- Four years developing Network security detectors
- Mainly focusing on AI-based security detectors

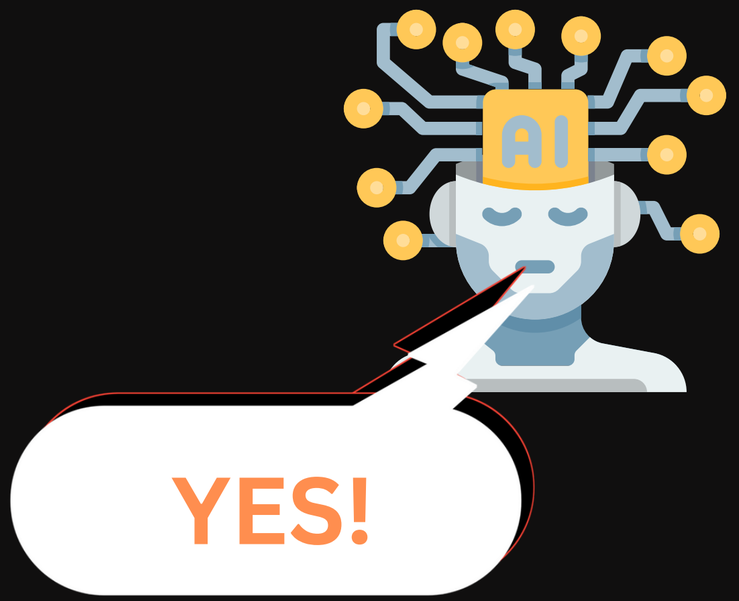
Accuracy of

99.9%

But it is hard to verify the output.

Are those hotdogs?

Are those hotdogs?

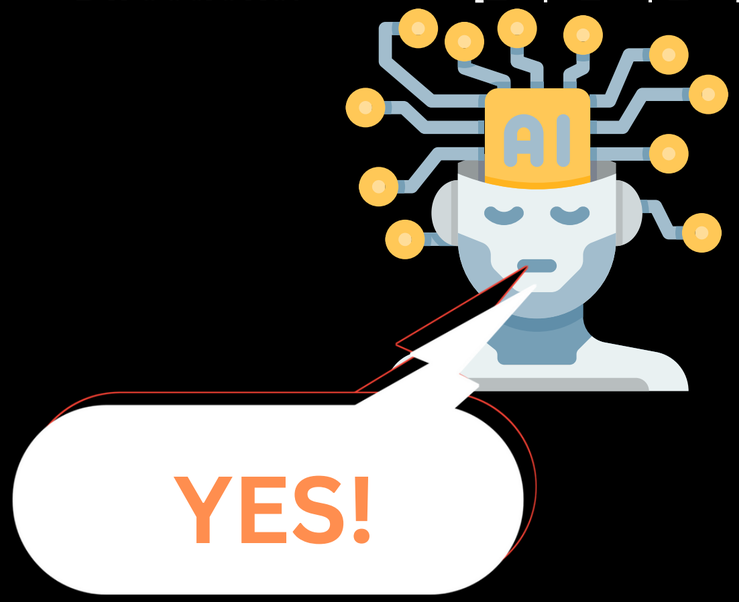


Is this malware?

Packet Sizes: [231|412|221|82|232|415|240|82|242|414|230|82|237|415|244|82|261|75|412|64],
Interpacket Times: [10 | 14 |443|13|14 | 53 |534|44|424|232|535|44|232| 1 | 13 |44|434|32|555|23]
Directions: [1 | -1 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 |1 | -1 | -1 | -1 | -1 | 1 |-1 | 1 | 1]

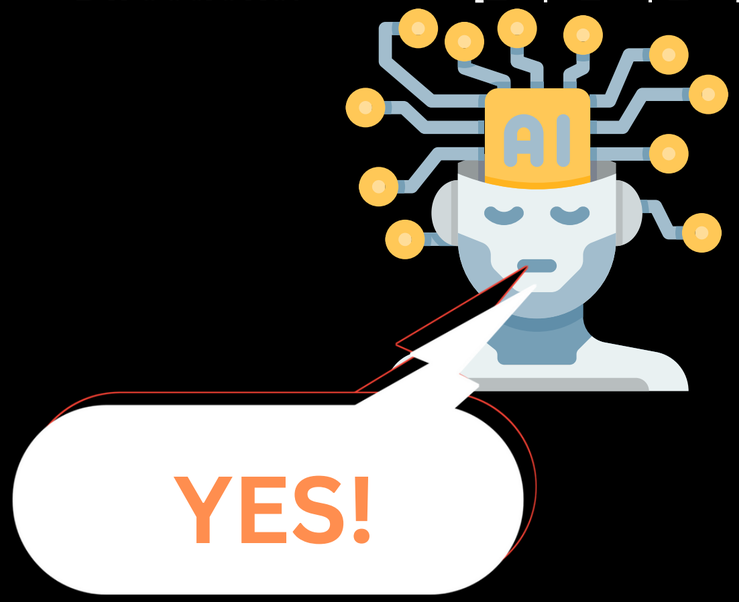
Is this malware?

Packet Sizes: [231|412|221|82|232|415|240|82|242|414|230|82|237|415|244|82|261|75|412|64]
Interpacket Times: [10 | 14 |443|13|14 | 53 |534|44|424|232|535|44|232| 1 | 13 |44|434|32|555|23]
Directions: [1 | -1 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | -1 | 1 | 1]



Is this malware?

Packet Sizes: [231|412|221|82|232|415|240|82|242|414|230|82|237|415|244|82|261|75|412|64]
Interpacket Times: [10 | 14 |443|13|14 | 53 |534|44|424|232|535|44|232| 1 | 13 |44|434|32|555|23]
Directions: [1 | -1 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | -1 | 1 | 1]



You can't quickly verify.

We must trust it...

99.9%

Quick Math...

10K Network telemetry records per second

Network of mid-sized company

99.9% Accuracy

Higher standard for network threat classification

1 false-detection per second

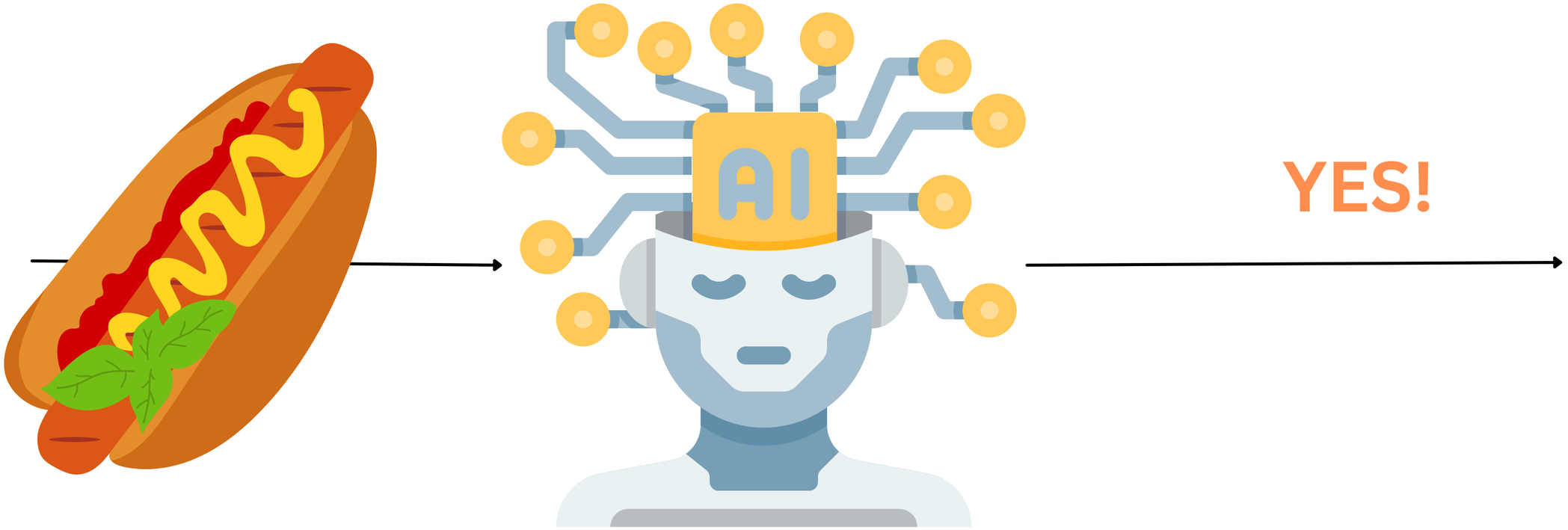
10K * 0.001

It is gonna be turned off very soon!

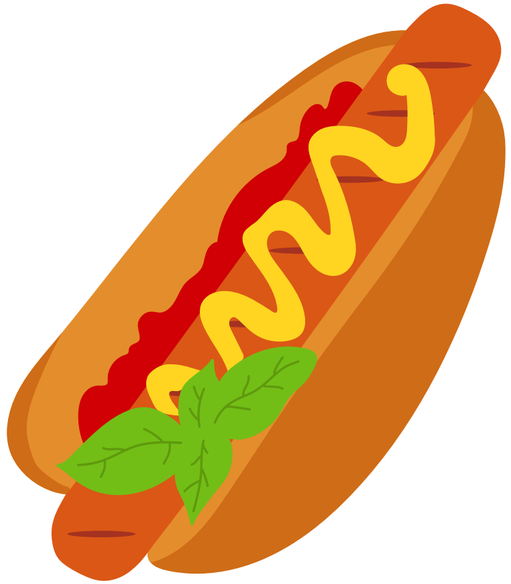
We need to distinguish wrong detections quickly...

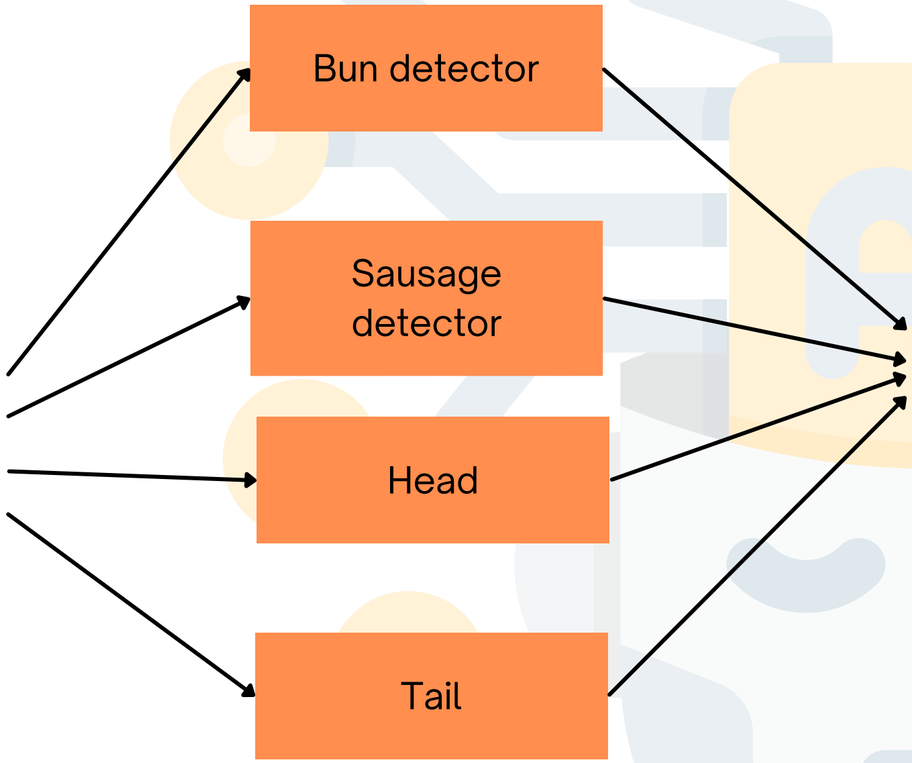
By adding explanations.

Monolithic AI



Dividing AI





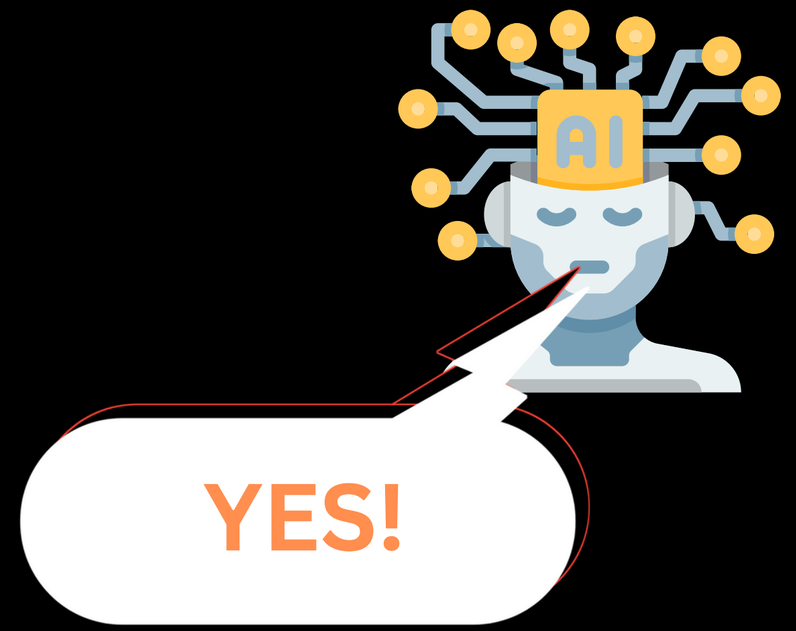
YES!

Because there is Bun and sausage

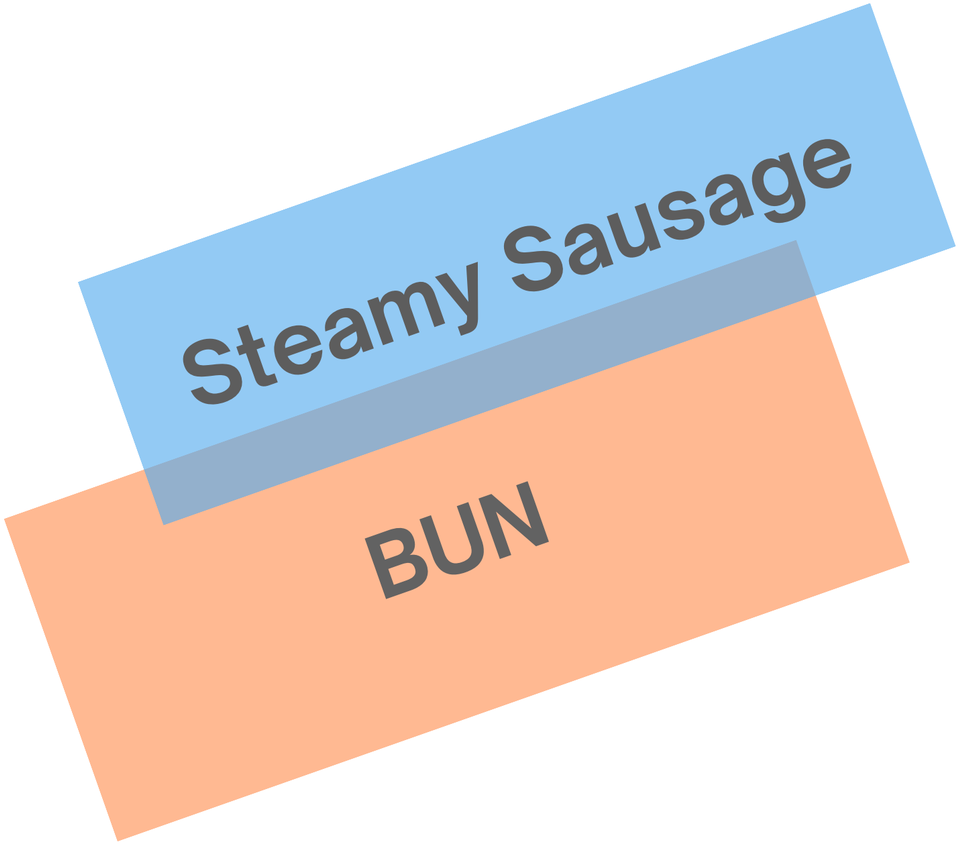
Is this a hotdog?

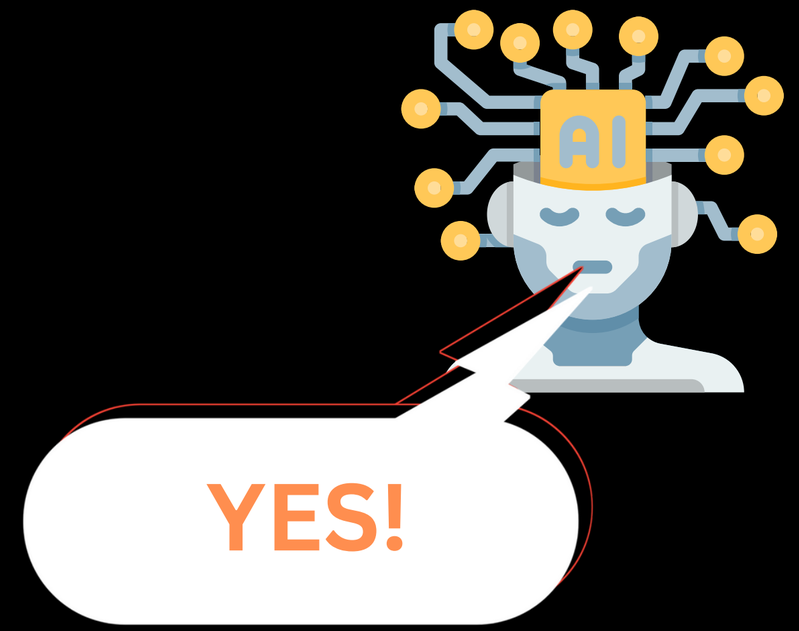
Is this a hotdog?

BUN

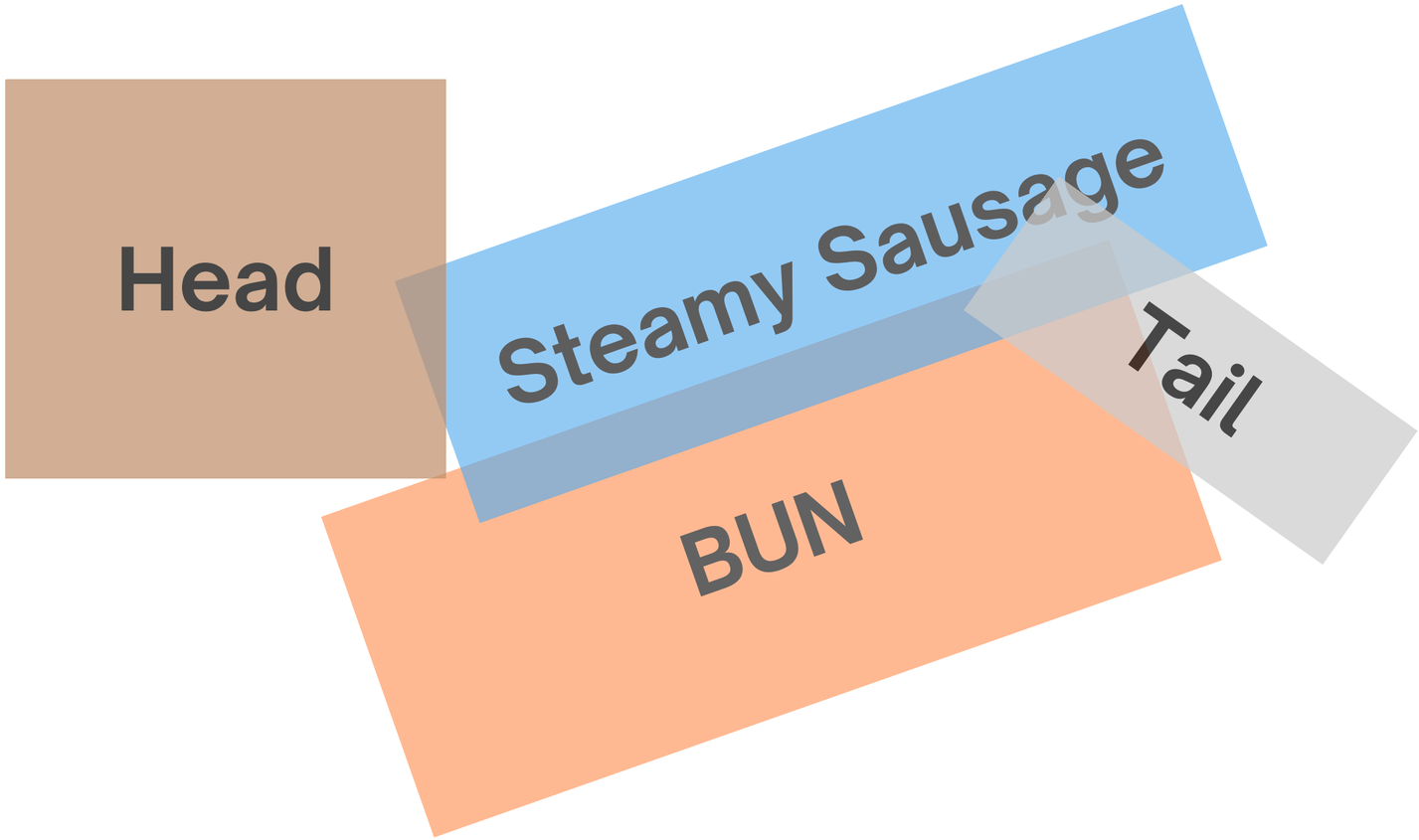


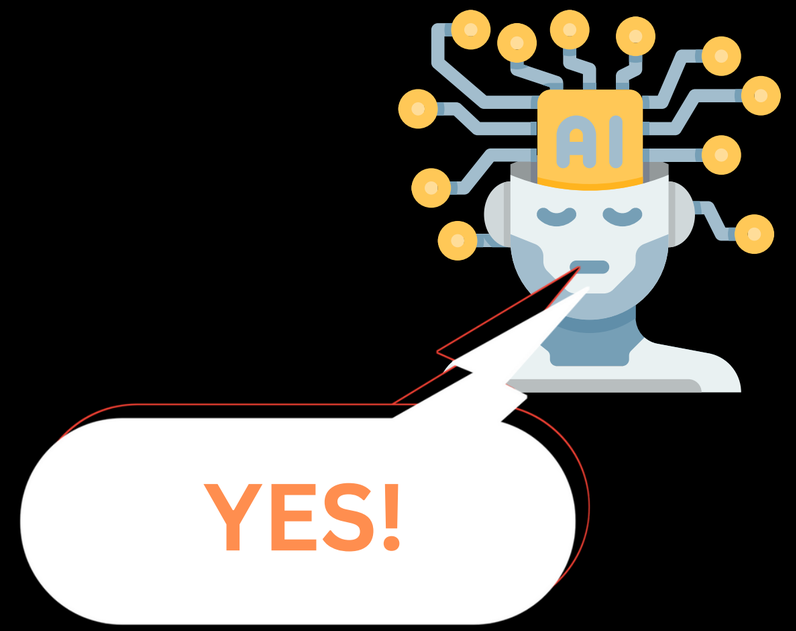
Is this a hotdog?





Is this a hotdog?





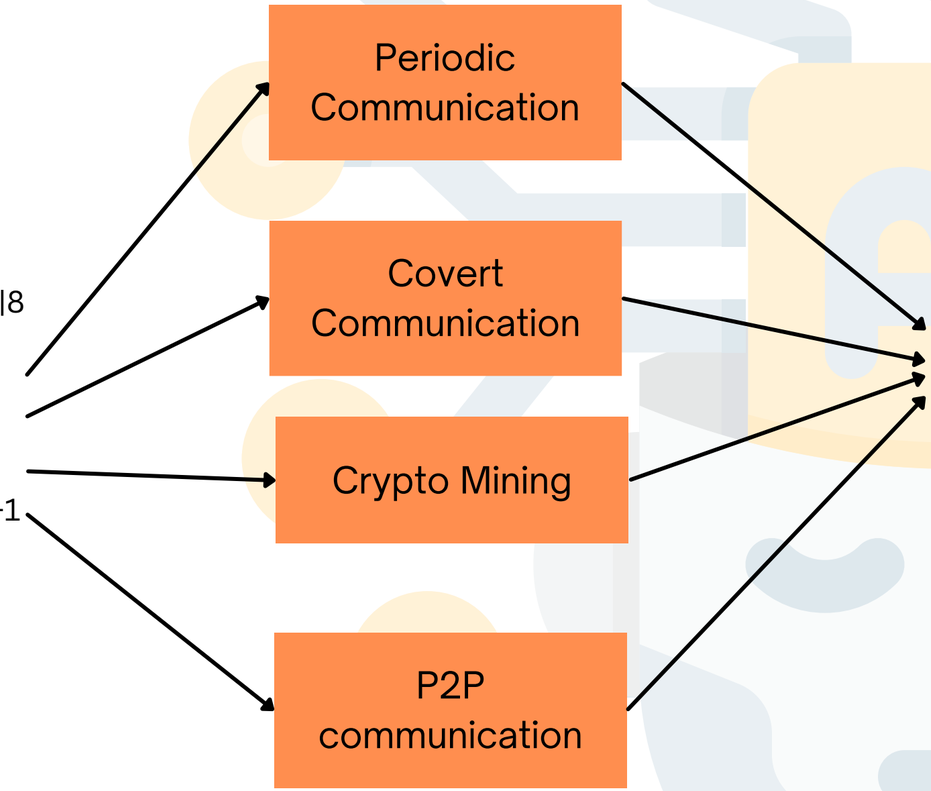
Something is really wrong!

Is this a hotdog?

Something is really wrong!

Dividing AI

[231|412|221|82|232|415|240|82|242|414|230|82|237|415|244|82|261|75|412|64]
[10 | 14 |443|13|14 | 53 |534|44|424|232|535|44|232| 1 | 13 |44|434|32|555|23]
[1 | -1 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 |1 | -1 | -1 | -1 | -1 | 1 |-1 | 1 | 1]



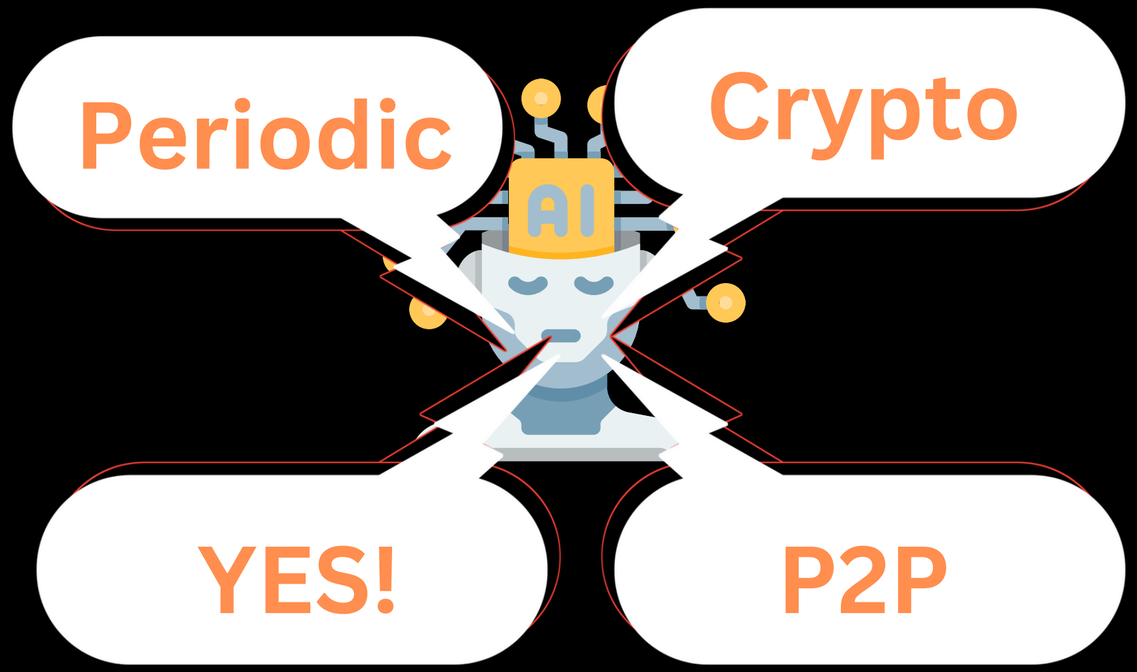
Combinator

YES!

Because I've found following properties...

Is this communication malicious?

Packet Sizes: [231|412|221|82|232|415|240|82|242|414|230|82|237|415|244|82|261|75|412|64]
Interpacket Times: [10 | 14 |443|13|14 | 53 |534|44|424|232|535|44|232| 1 | 13 |44|434|32|555|23]
Directions: [1 | -1 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | -1 | 1 | 1]



The benefits of divided AI

1. We can quickly identify misclassifications
2. Less prone to design errors
3. Even non-AI expert can understand the explanations

You can contact me at

karel.hynek@cesnet.cz

Karel Hynek

Network Security Research and Development
CESNET a.l.e.