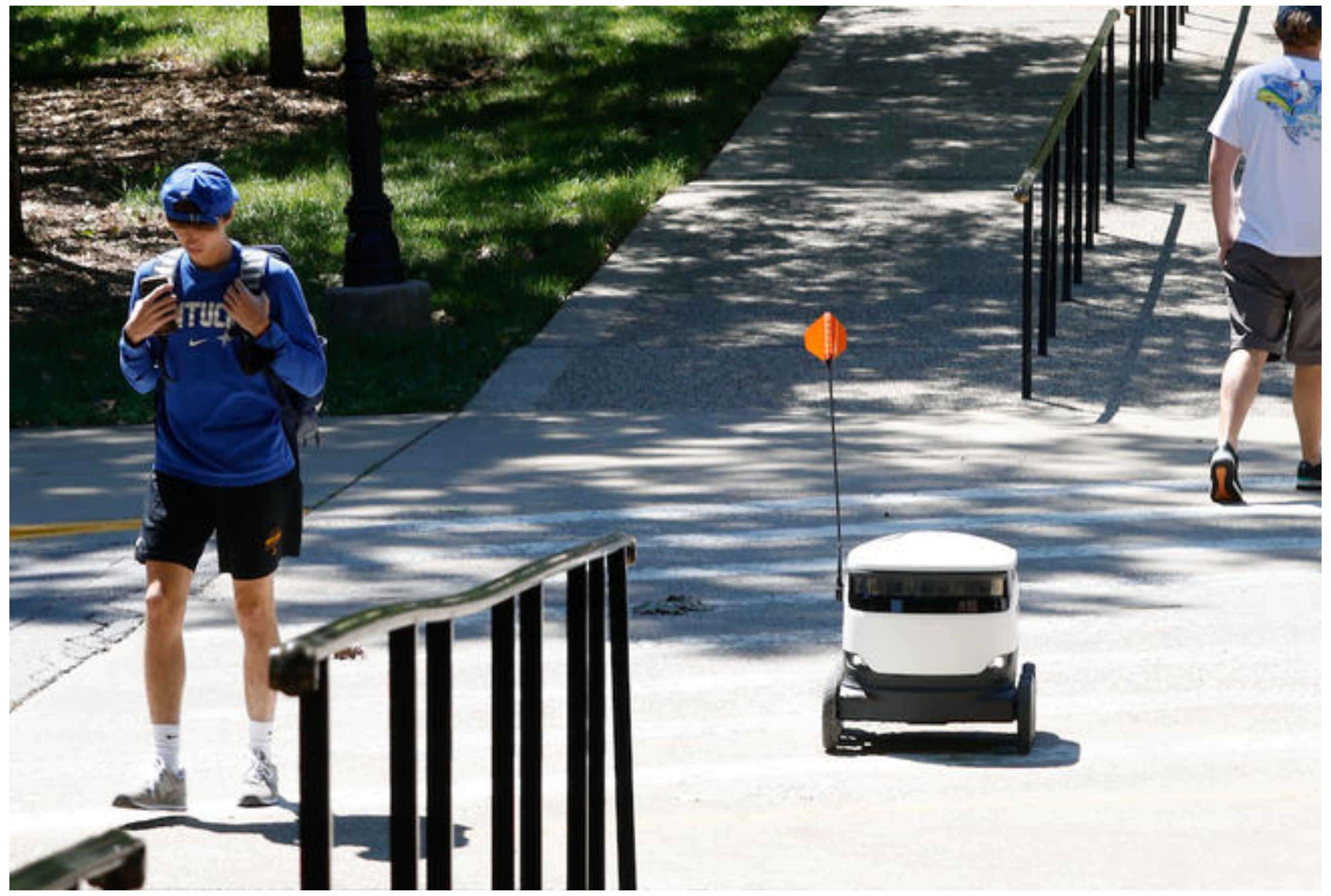
GONGBO "TONY" LIANG, PHD (GLIANG@TAMUSA.EDU)

ADDRESSING TRUST AND SAFETY CHALLENGES IN NEURAL NETWORK-POWERED MODERN AI:

A call for broader awareness and action

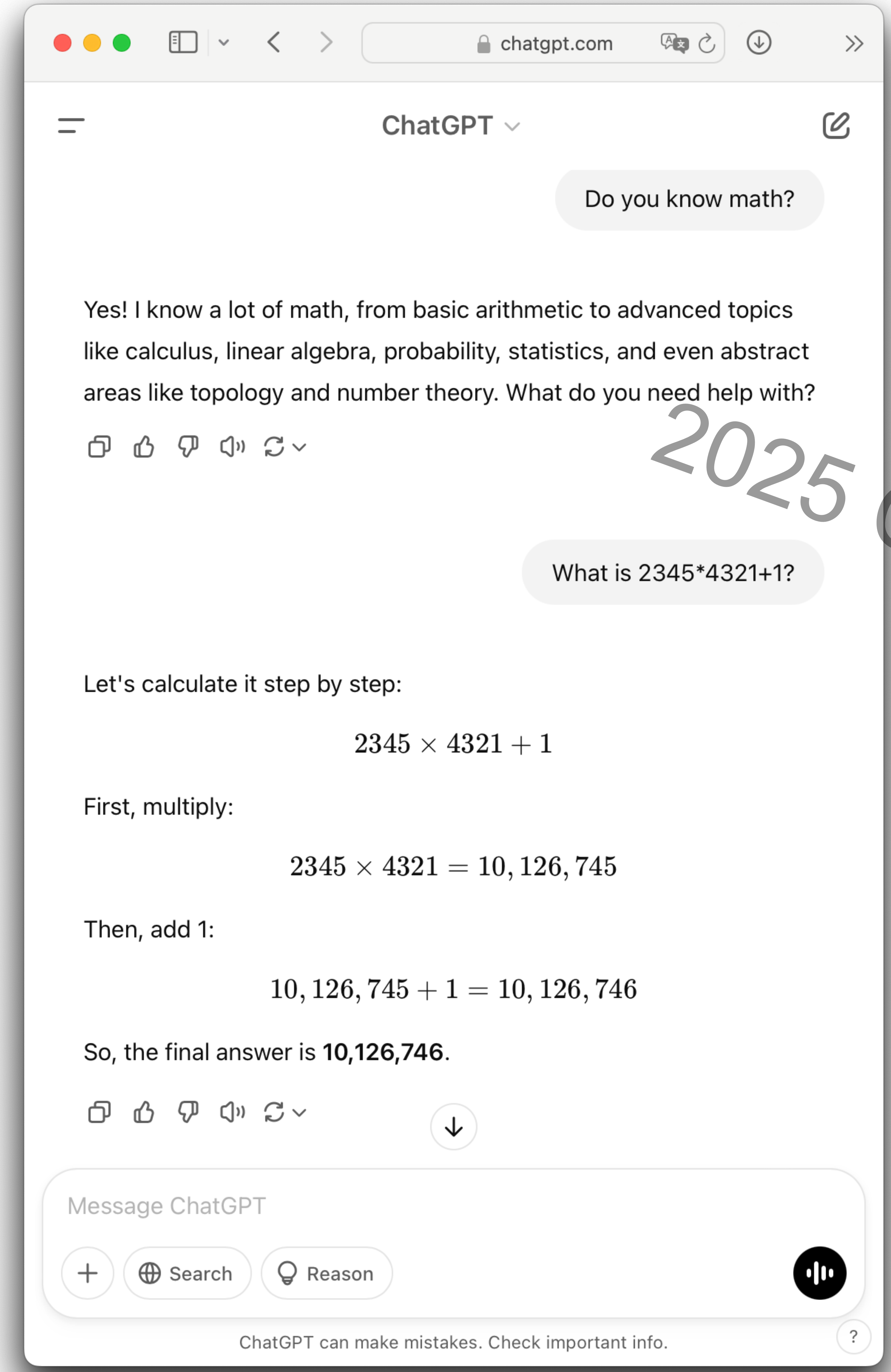
TEXAS A&M UNIVERSITY
SAN ANTONIO

SCI-FI AI VS REALITY

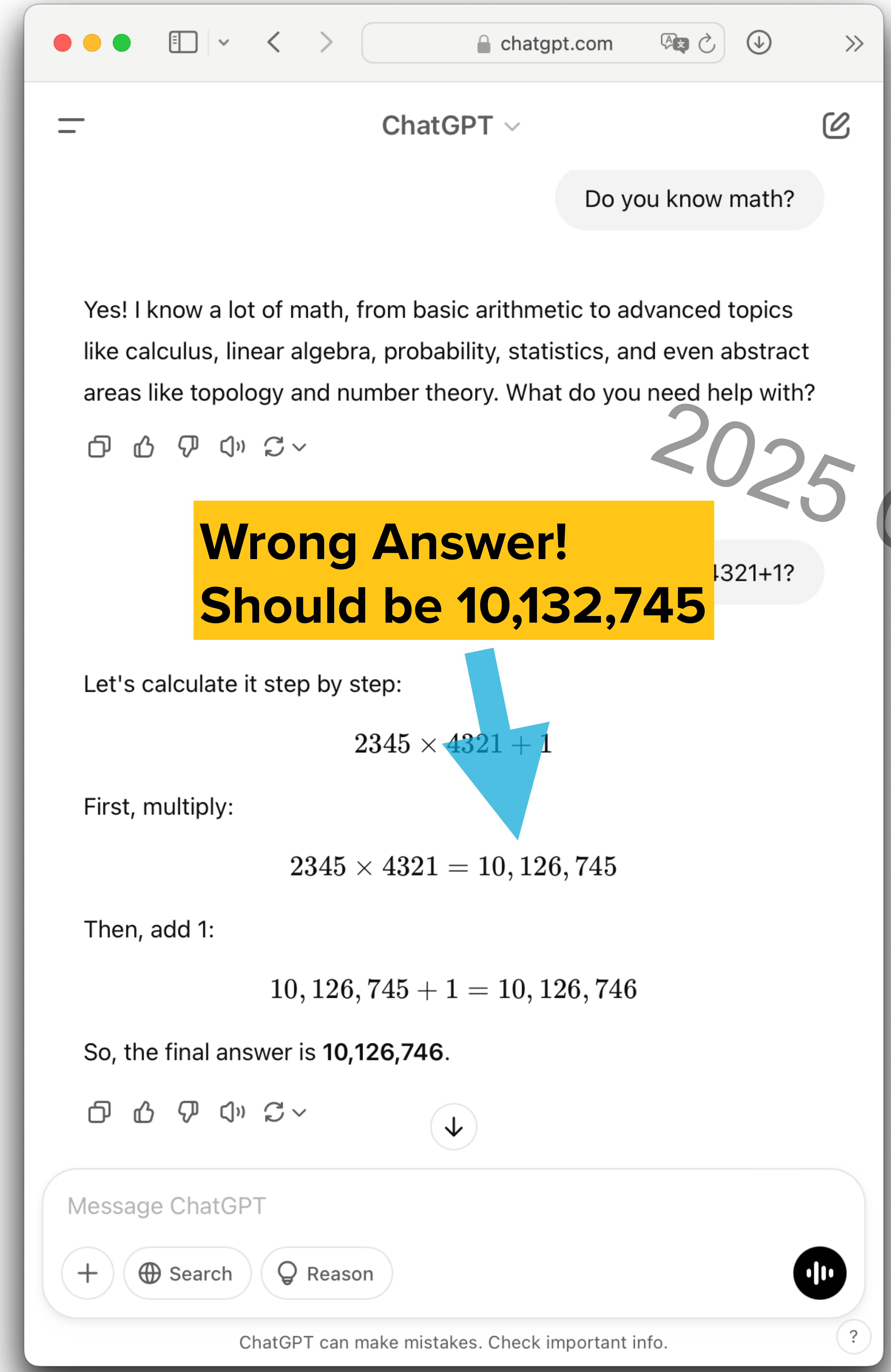



"AI IS THE NEW ELECTRICITY"
Dr. Andrew Ng
https://scet.berkeley.edu/ai-is-the-new-electricity-insights-from-dr-andrew-ng/

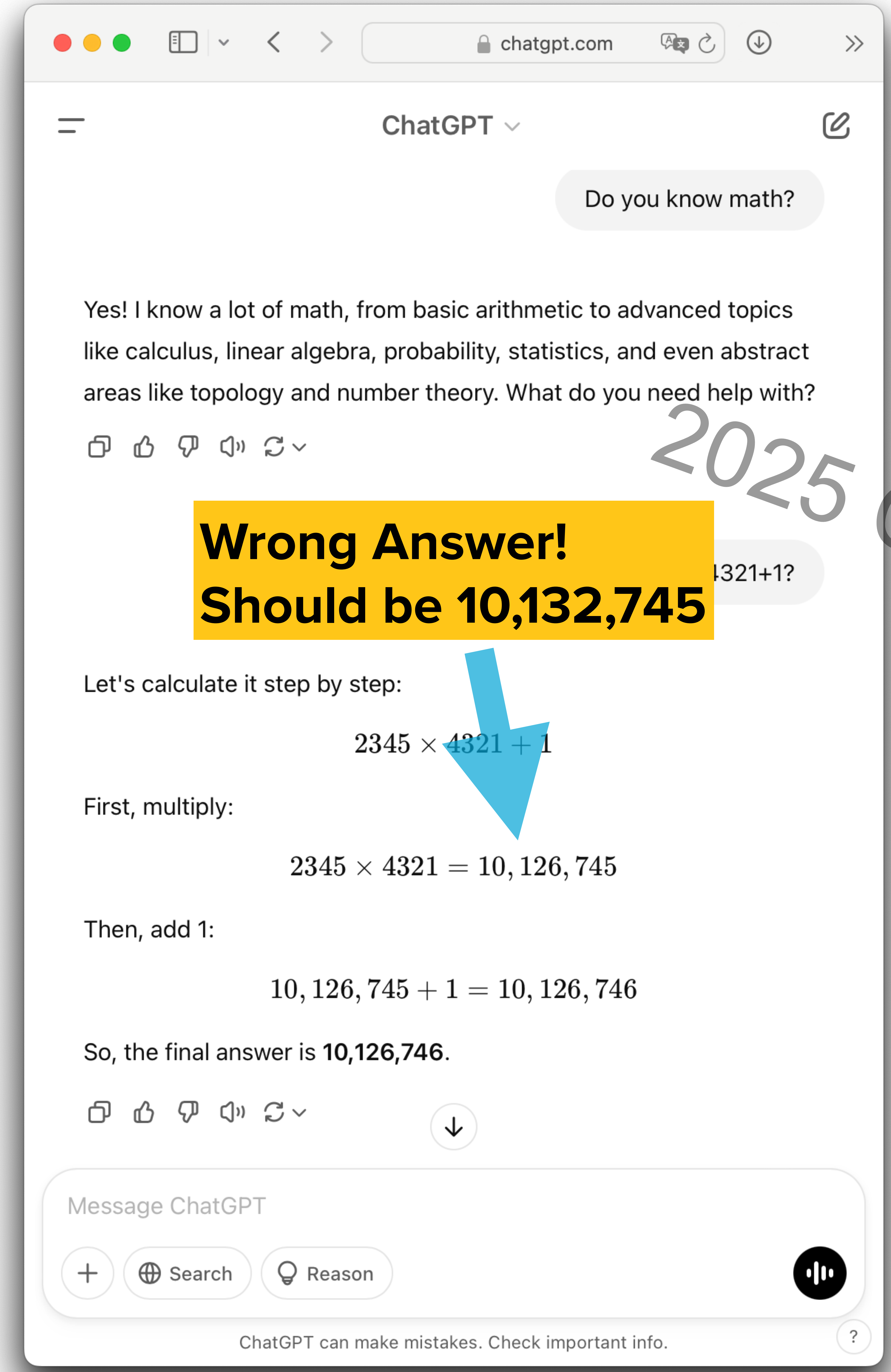
CAN WE TRUST AI?

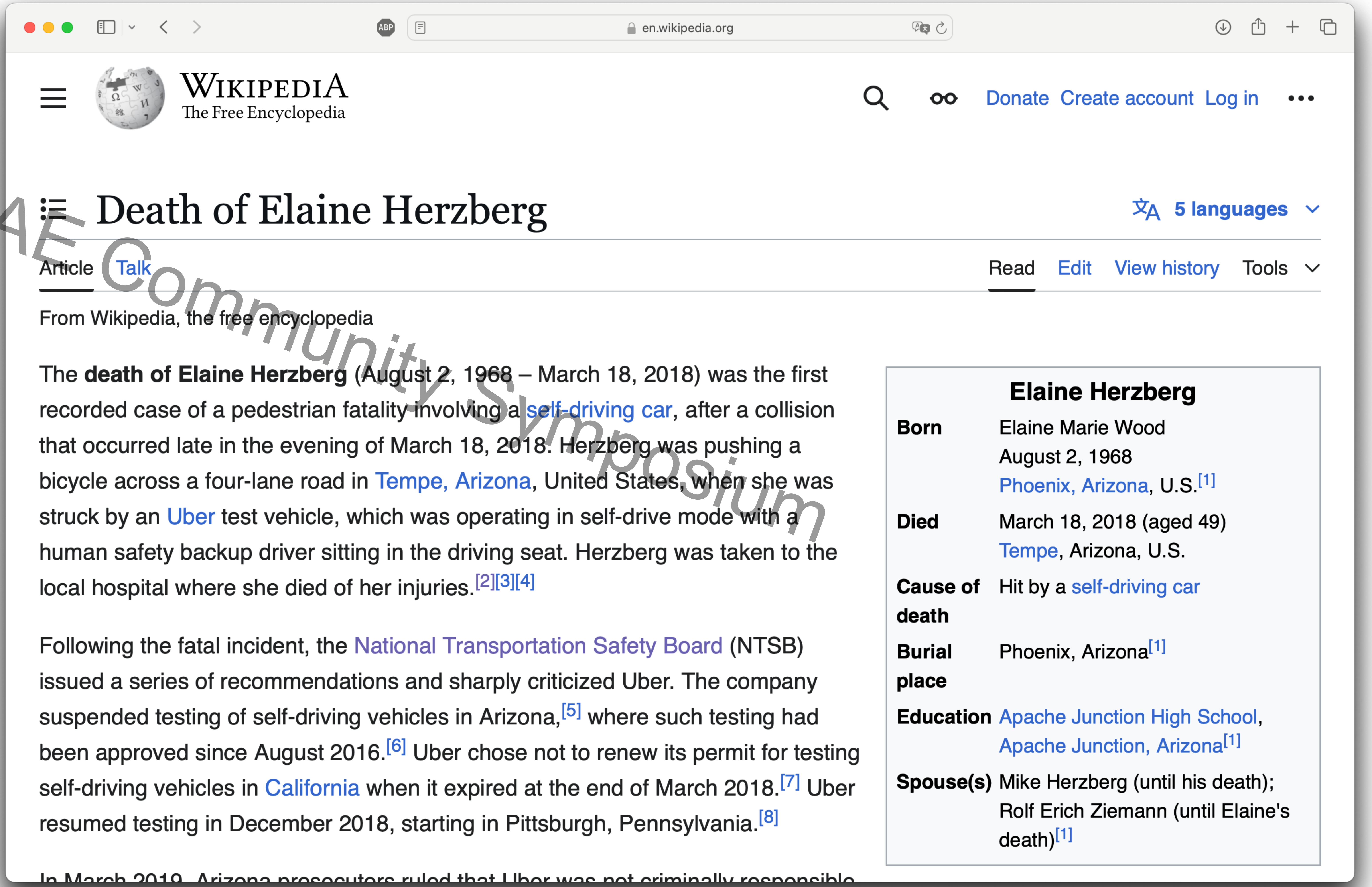


CAN WE TRUST AI?

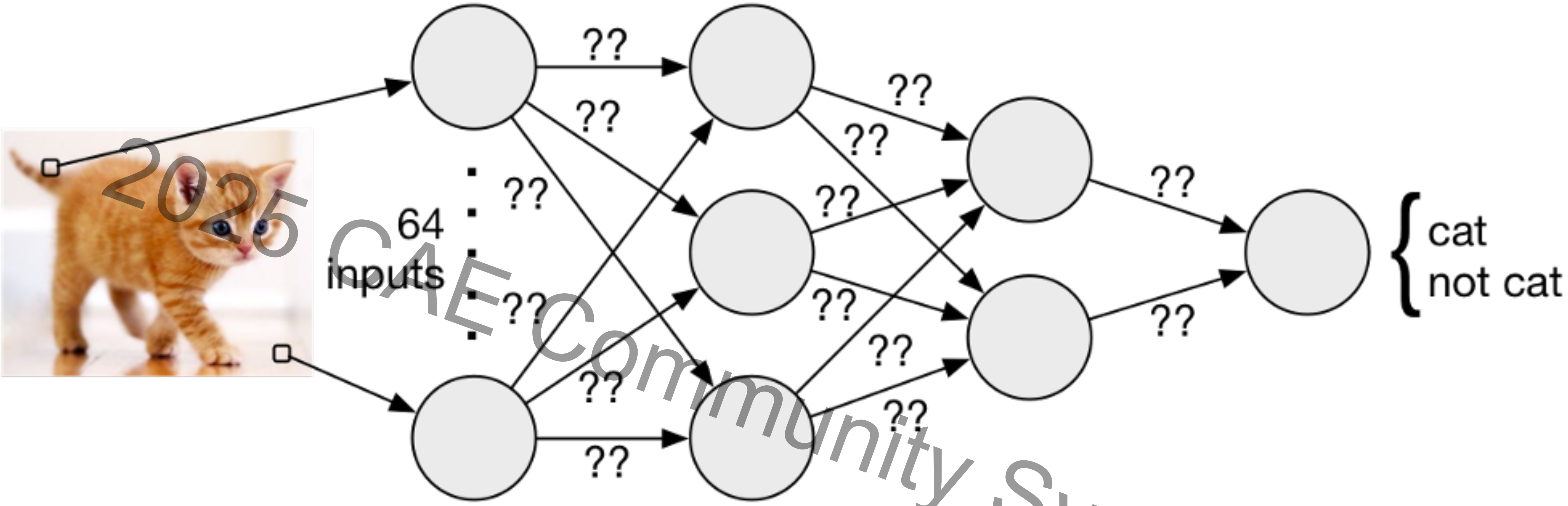


CAN WE TRUST AI?

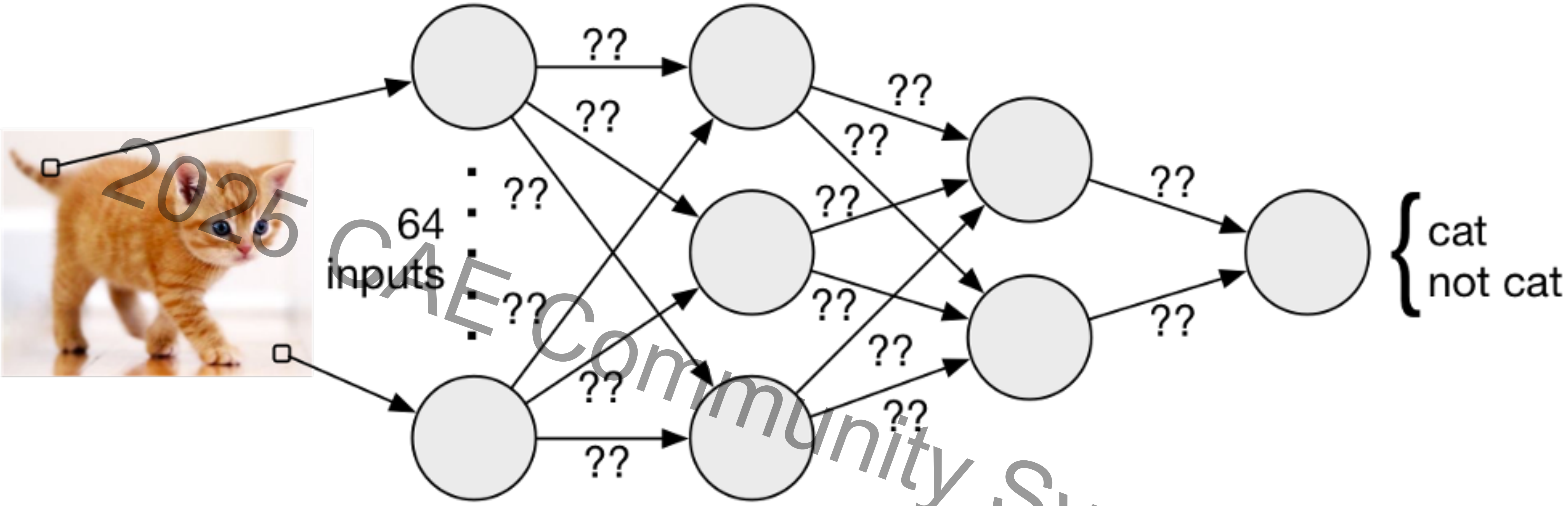


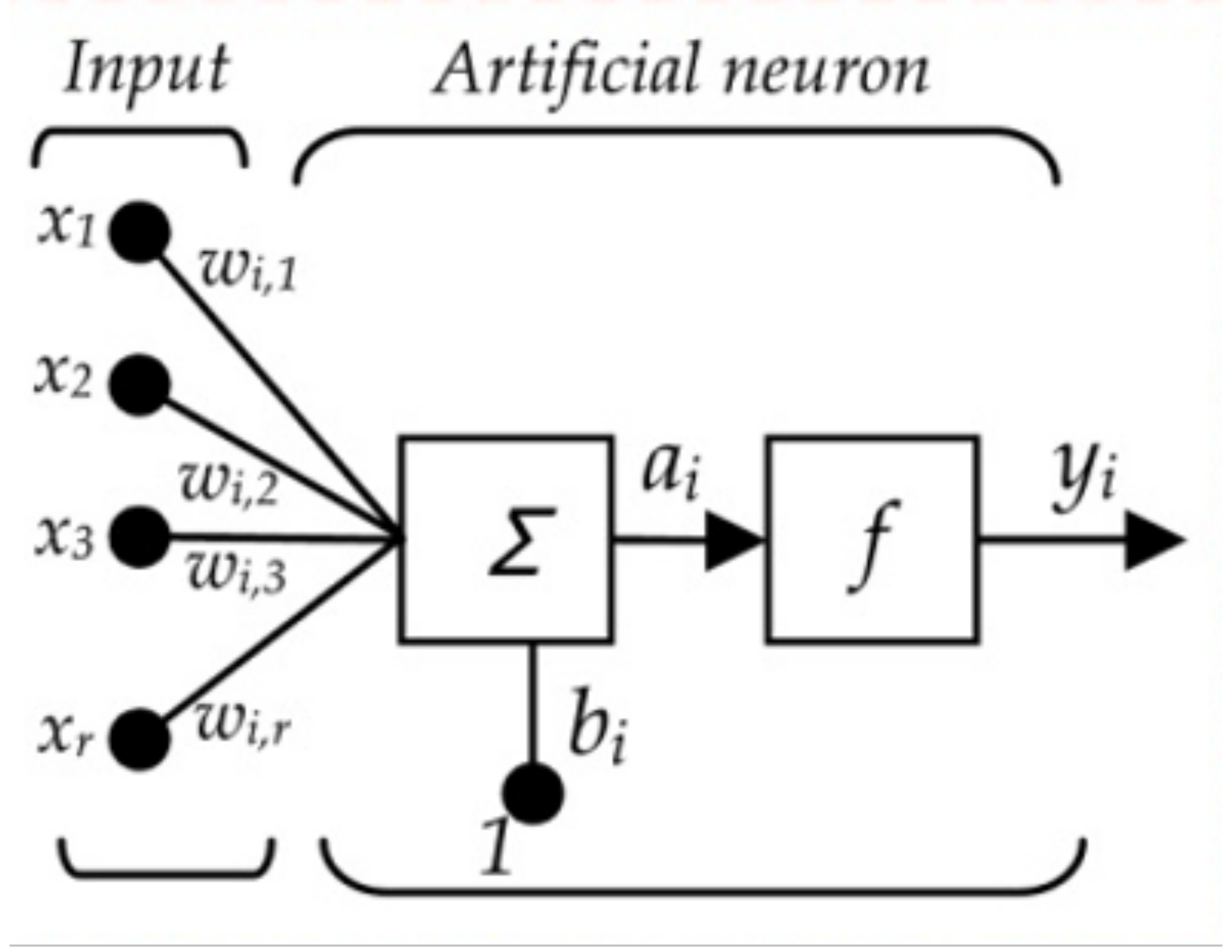


NEURAL NETWORK REPRESENTATION

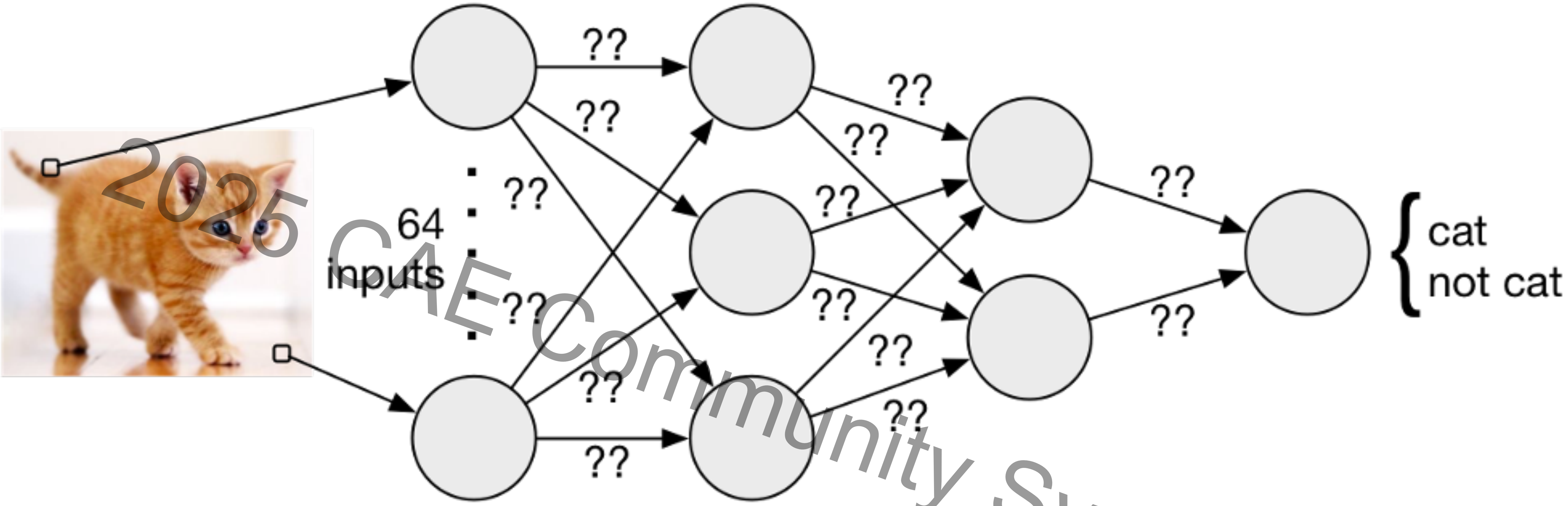


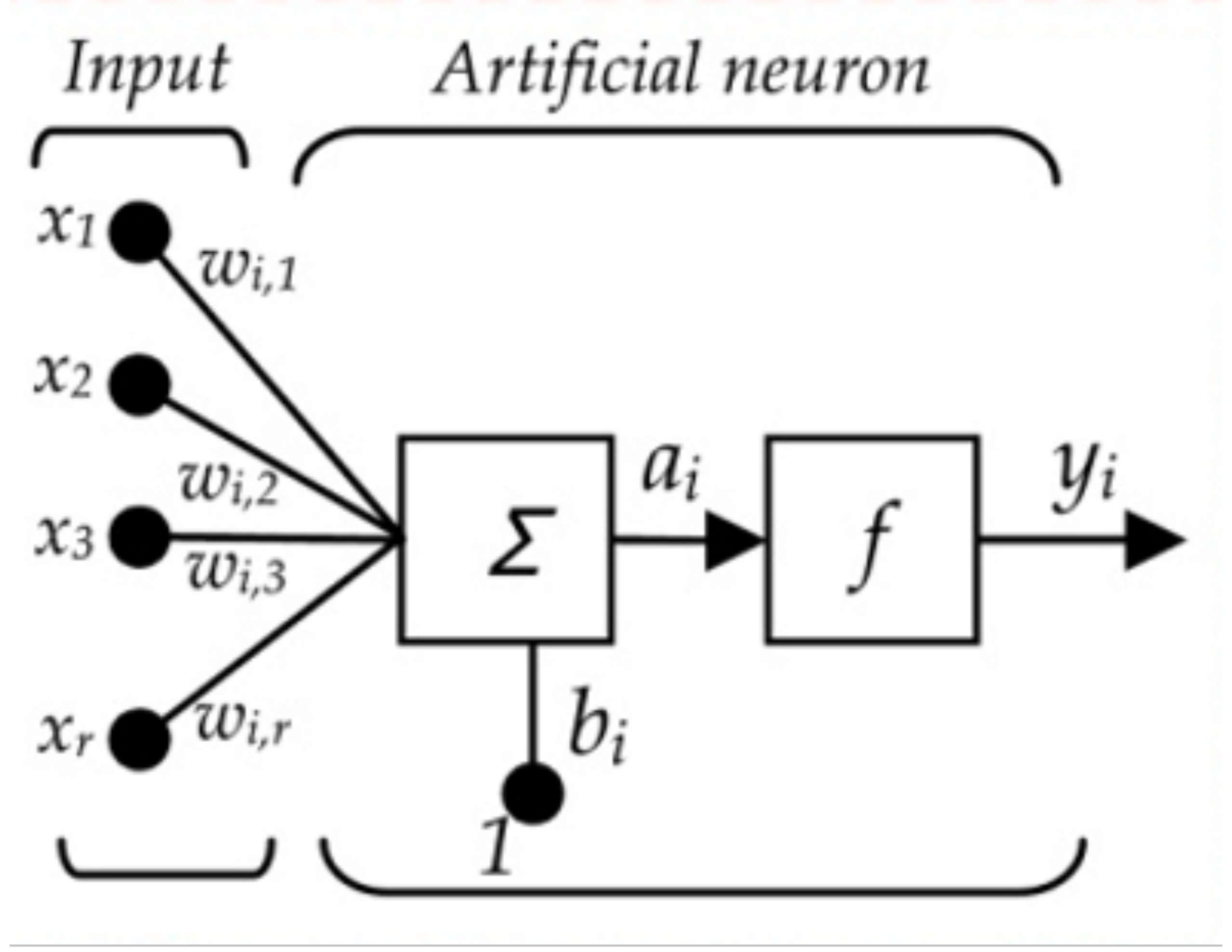
NEURAL NETWORK REPRESENTATION





NEURAL NETWORK REPRESENTATION





$$h_W(X) = f(W^T X)$$

$$f(z) = \frac{1}{1 + e^{-z}}$$

NEURAL NETWORK UNCERTAINTY ESTIMATION

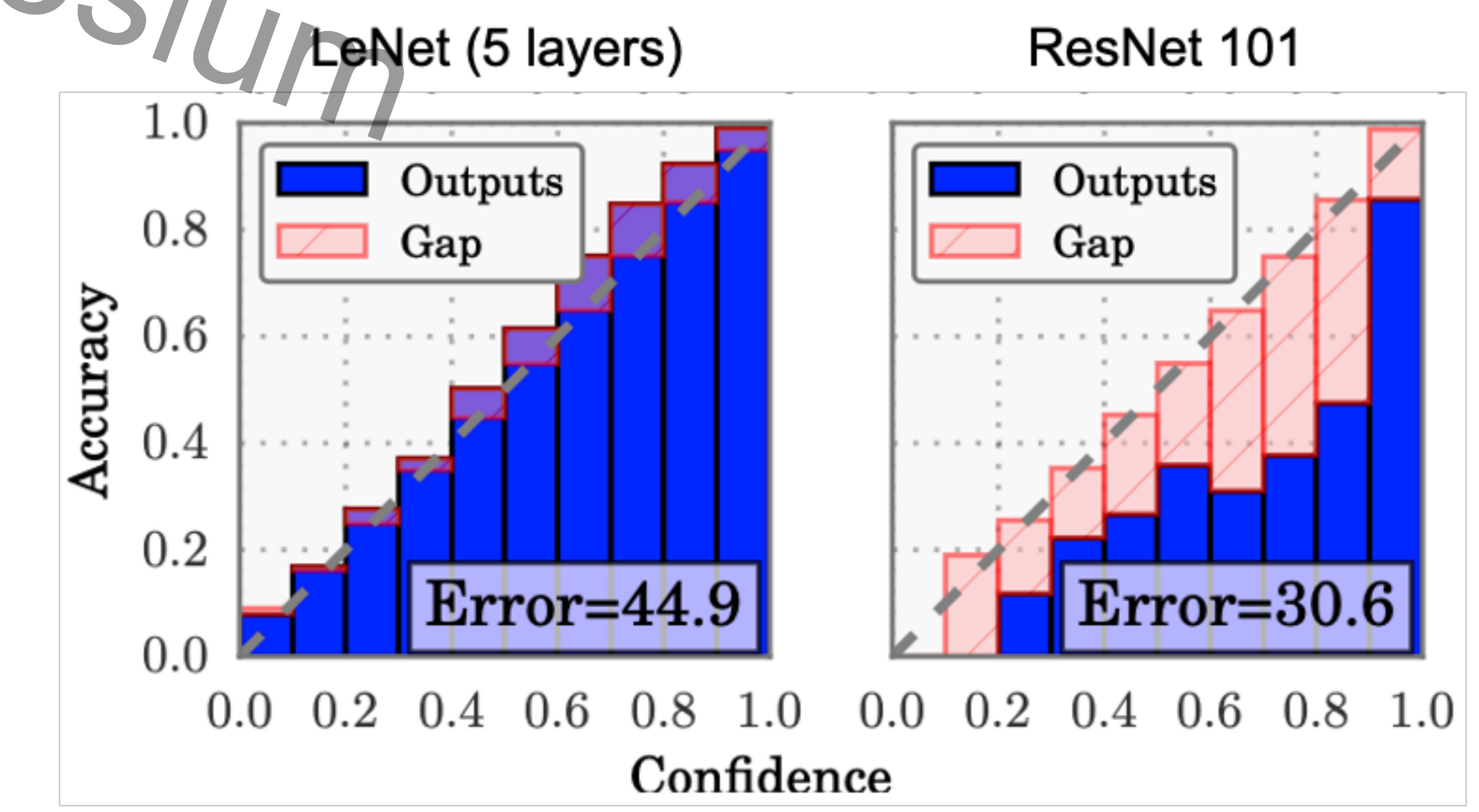
- Let
  - $x \in X$: input
  - $y \in Y = \{1, \ldots, k\}$: target label
  - $h()$: a neural network
    - $h(x) = (\hat{y}, \hat{p})$
    - $\hat{y}$: predicted label
    - $\hat{p}$: predicted probability/confidence

- Ideally: $\mathbb{P}(\hat{y} = y | \hat{p} = p) = P, \forall P \in [0,1]$
  - P is the true data distribution

- In reality: $\mathbb{P}(\hat{y} = y | \hat{p} = p) \neq P$

Example I:
- For any binary classification tasks, given 100 predictions with an average confidence of 0.95
- We would expect that around 95 correct predictions

Example II:
- For any multi-class classification tasks, given 100 predictions with 95% accuracy
- We would expect that average confidence of the correct prediction is around 0.95



Performance on CIFAR-100

Guo, Chuan, et al. "On calibration of modern neural networks." In *ICML*, 2017.

TRUSTWORTHY ISSUES WITH NEURAL NETWORKS

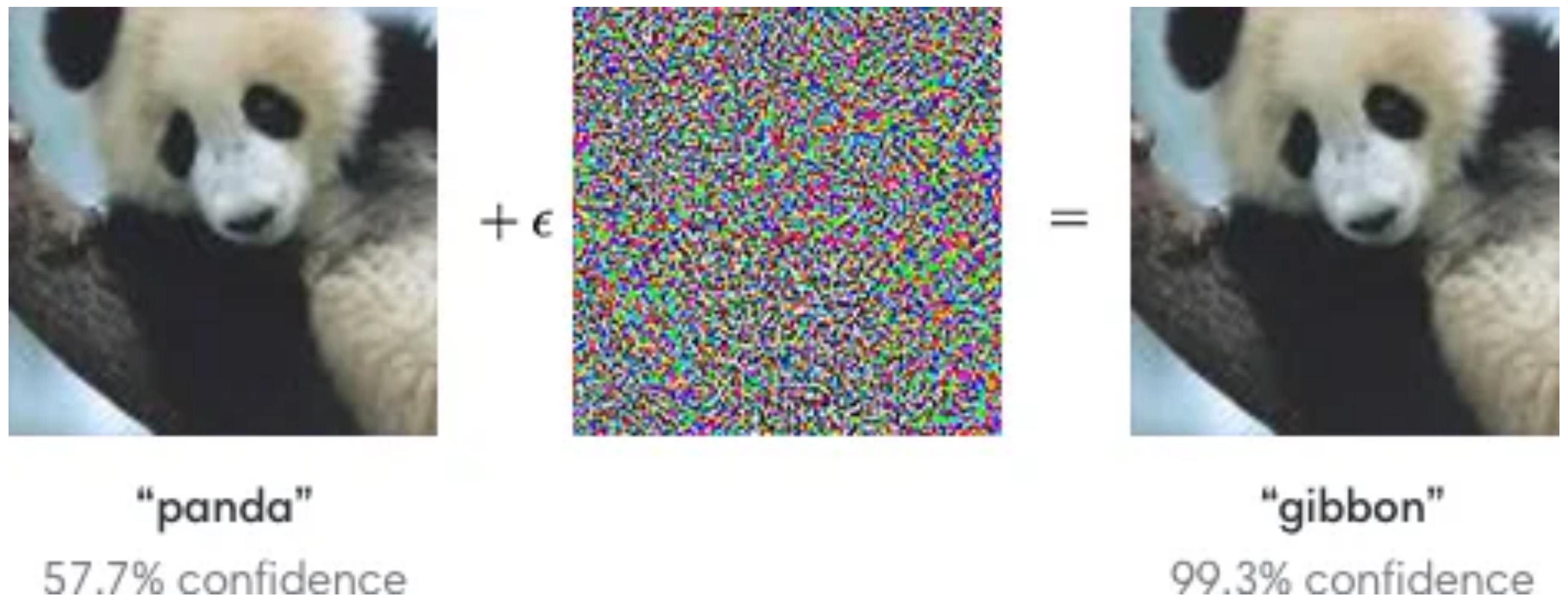
NOT AN EXHAUSTIVE LIST

- **Adversarial attacks:** attempts to trick predictive models into making incorrect predictions or decisions

TRUSTWORTHY ISSUES WITH NEURAL NETWORKS

NOT AN EXHAUSTIVE LIST

- **Adversarial attacks:** attempts to trick predictive models into making incorrect predictions or decisions

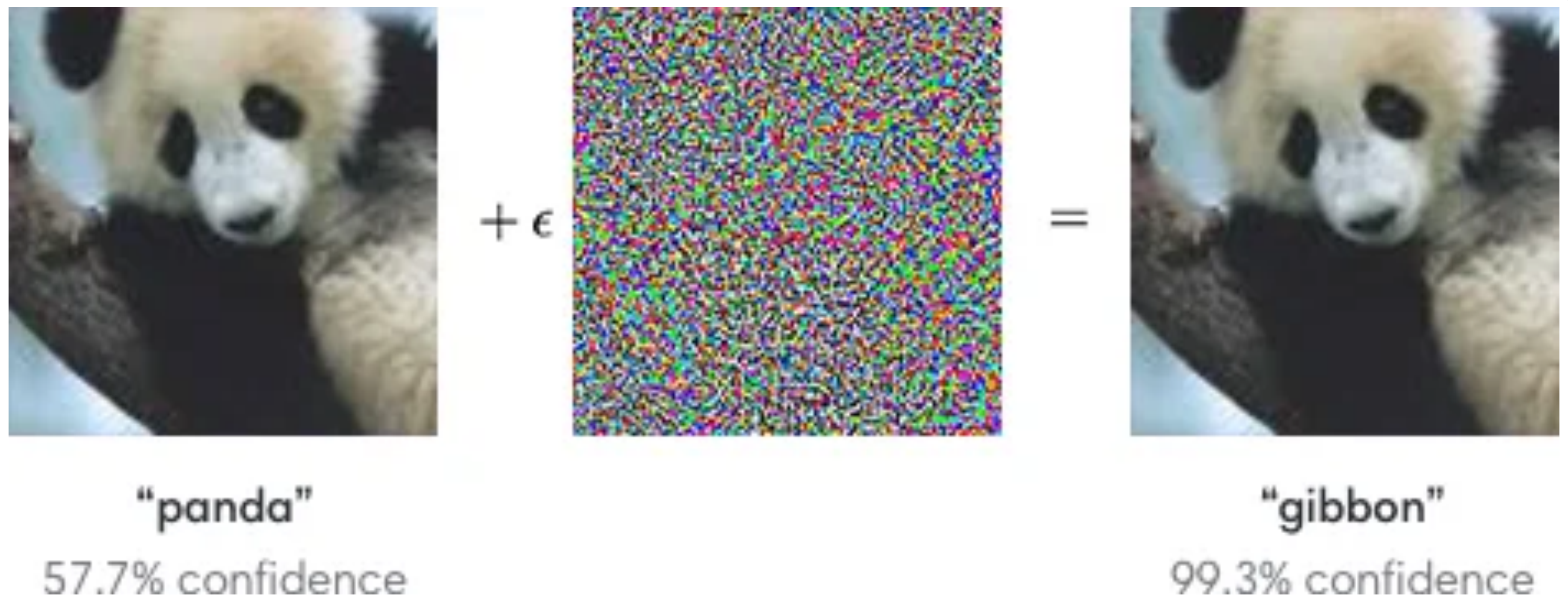


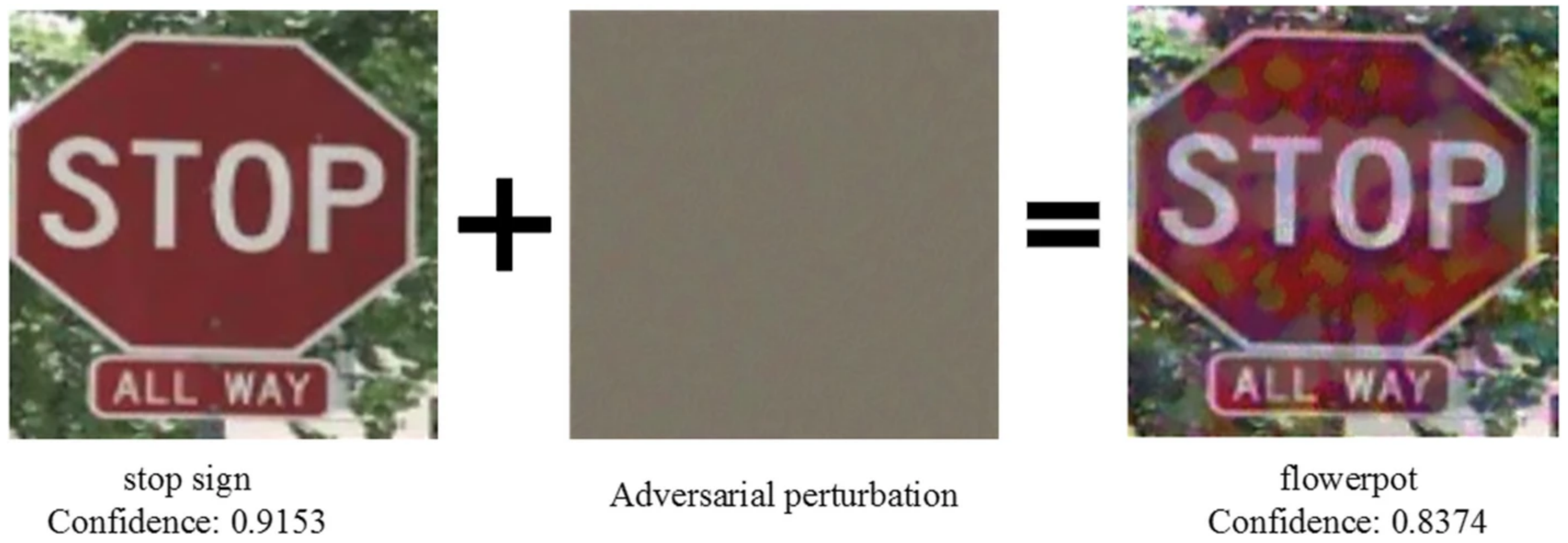
Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

TRUSTWORTHY ISSUES WITH NEURAL NETWORKS

NOT AN EXHAUSTIVE LIST

- **Adversarial attacks:** attempts to trick predictive models into making incorrect predictions or decisions





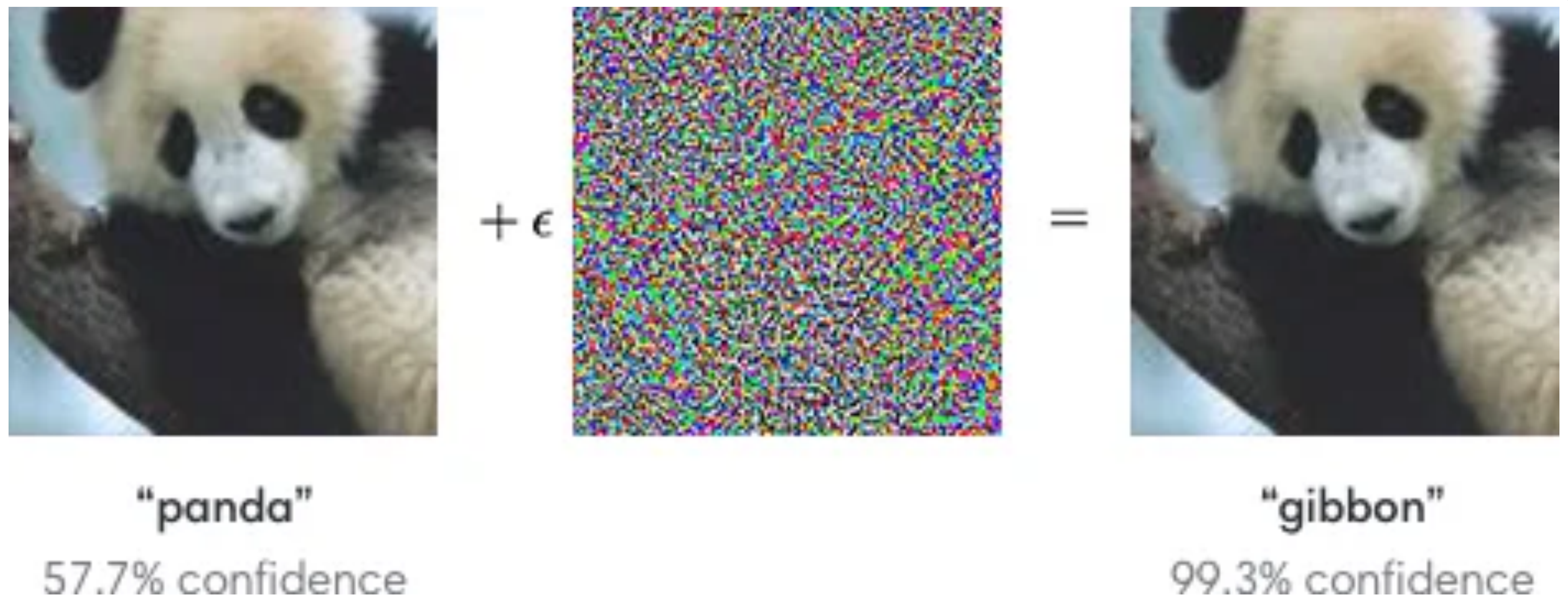
Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

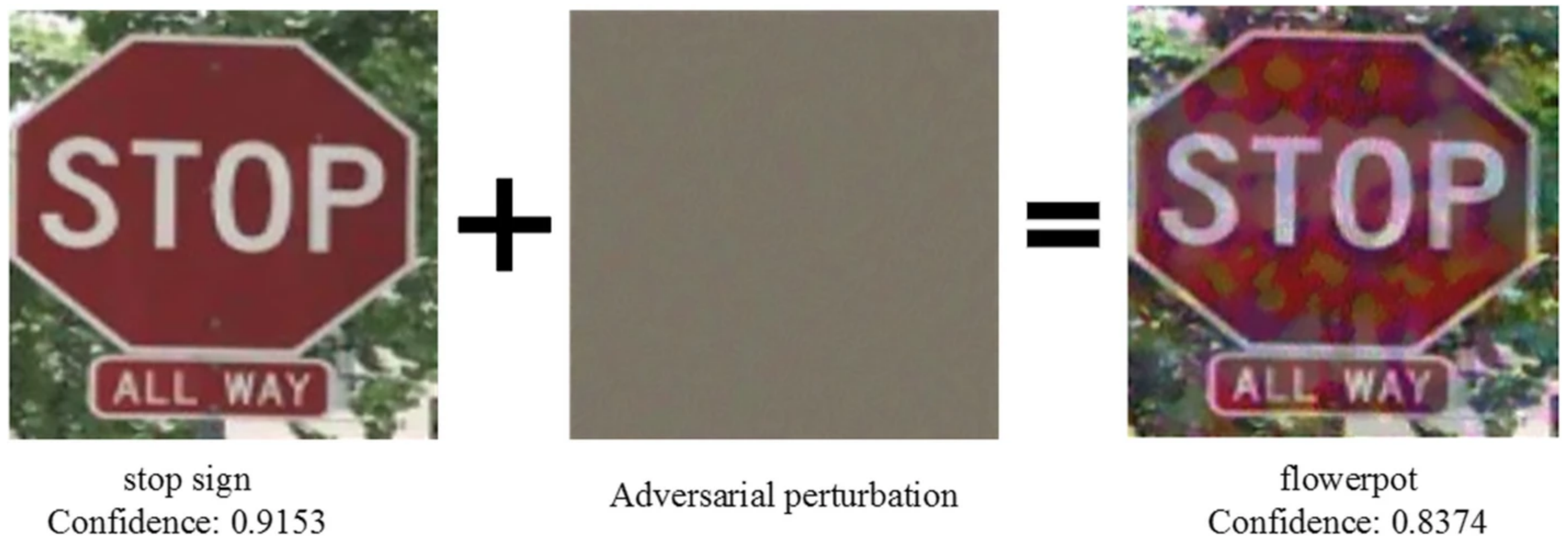
Wu, Fei, et al. "Defense against adversarial attacks in traffic sign images identification based on 5G." EURASIP Journal on Wireless Communications and Networking 2020 (2020): 1-15.

TRUSTWORTHY ISSUES WITH NEURAL NETWORKS

NOT AN EXHAUSTIVE LIST

- **Adversarial attacks:** attempts to trick predictive models into making incorrect predictions or decisions





Original
Perfect performance by the actor → **Positive (99%)**

Adversarial
Spotless performance by the actor → **Negative (100%)**

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

Wu, Fei, et al. "Defense against adversarial attacks in traffic sign images identification based on 5G." EURASIP Journal on Wireless Communications and Networking 2020 (2020): 1-15.

Morris, John X. et al. "TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP." Conference on Empirical Methods in Natural Language Processing (2020).

TRUSTWORTHY ISSUES WITH NEURAL NETWORKS

NOT AN EXHAUSTIVE LIST

- **Adversarial attacks:** attempts to trick predictive models into making incorrect predictions or decisions

Mutation Operator	Mutated Sentence
Random misspelling words	**Plz** share and like the video
Random deleting articles	Please share and like ~~**the**~~ video
Random replacing a word with another one	Please **roar** and like the video
Random replacing a word with its synonym	Please **disseminate** and like the video
Random replacing a word with its antonym	Please share and **hate** the video
Random replacing "a" with "α"	Please **shαre** and like the video
Random replacing "e" with "ϵ"	Please share and **likϵ** the **vidϵo**

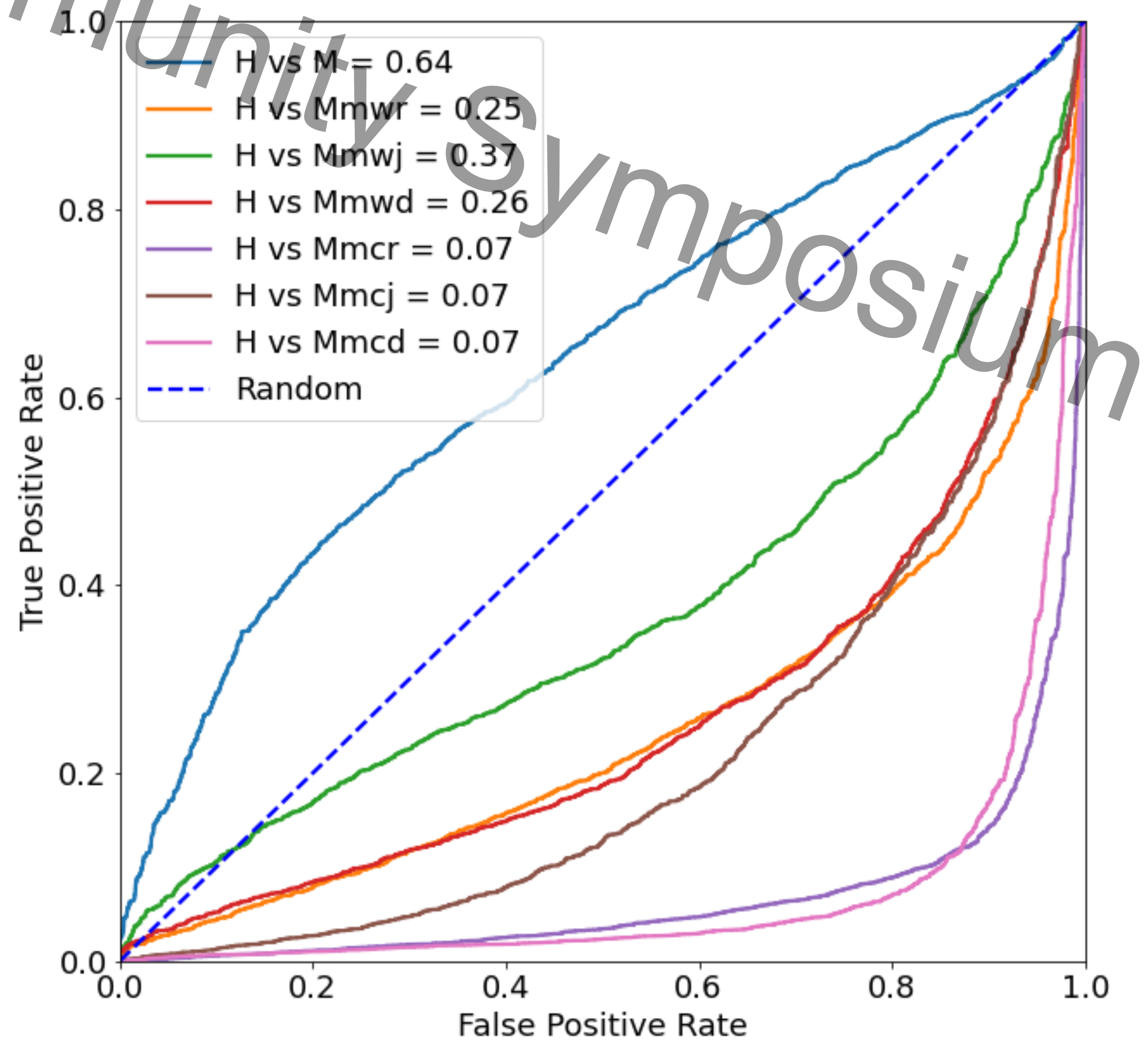
Liang, Gongbo, et al. "Enhancing neural text detector robustness with μ attacking and rr-training." Electronics 12.8 (2023): 1948.

TRUSTWORTHY ISSUES WITH NEURAL NETWORKS

NOT AN EXHAUSTIVE LIST

- **Adversarial attacks:** attempts to trick predictive models into making incorrect predictions or decisions

Mutation Operator	Mutated Sentence
Random misspelling words	**Plz** share and like the video
Random deleting articles	Please share and like ~~**the**~~ video
Random replacing a word with another one	Please **roar** and like the video
Random replacing a word with its synonym	Please **disseminate** and like the video
Random replacing a word with its antonym	Please share and **hate** the video
Random replacing “a" with “α"	Please **shαre** and like the video
Random replacing “e" with “ϵ"	Please share and **likϵ** the **vidϵo**



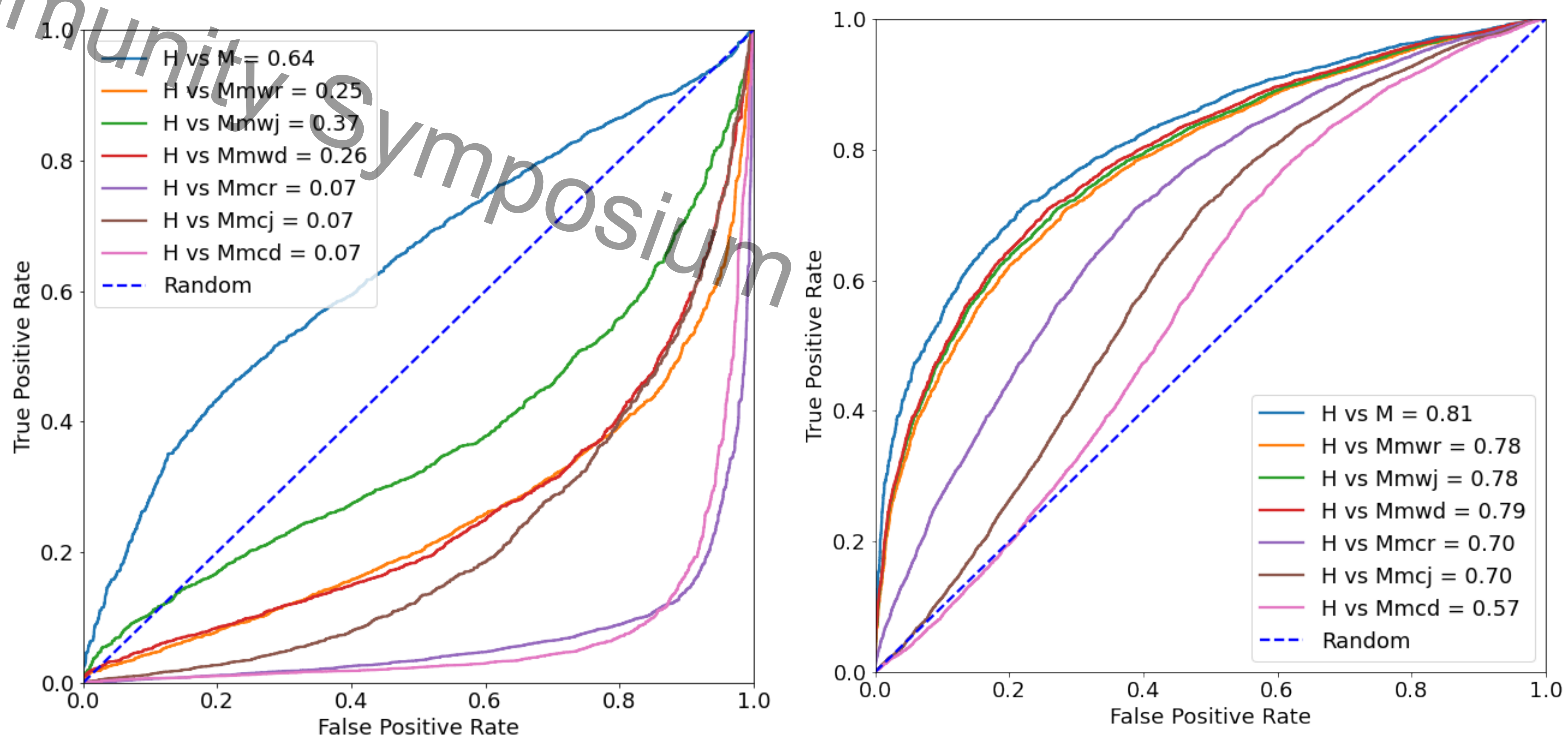
Liang, Gongbo, et al. "Enhancing neural text detector robustness with μ attacking and rr-training." Electronics 12.8 (2023): 1948.

TRUSTWORTHY ISSUES WITH NEURAL NETWORKS

NOT AN EXHAUSTIVE LIST

- **Adversarial attacks:** attempts to trick predictive models into making incorrect predictions or decisions

Mutation Operator	Mutated Sentence
Random misspelling words	**Plz** share and like the video
Random deleting articles	Please share and like **~~the~~** video
Random replacing a word with another one	Please **roar** and like the video
Random replacing a word with its synonym	Please **disseminate** and like the video
Random replacing a word with its antonym	Please share and **hate** the video
Random replacing "a" with "α"	Please **shαre** and like the video
Random replacing "e" with "ϵ"	Please share and **likϵ** the **vidϵo**



Liang, Gongbo, et al. "Enhancing neural text detector robustness with μ attacking and rr-training." Electronics 12.8 (2023): 1948.

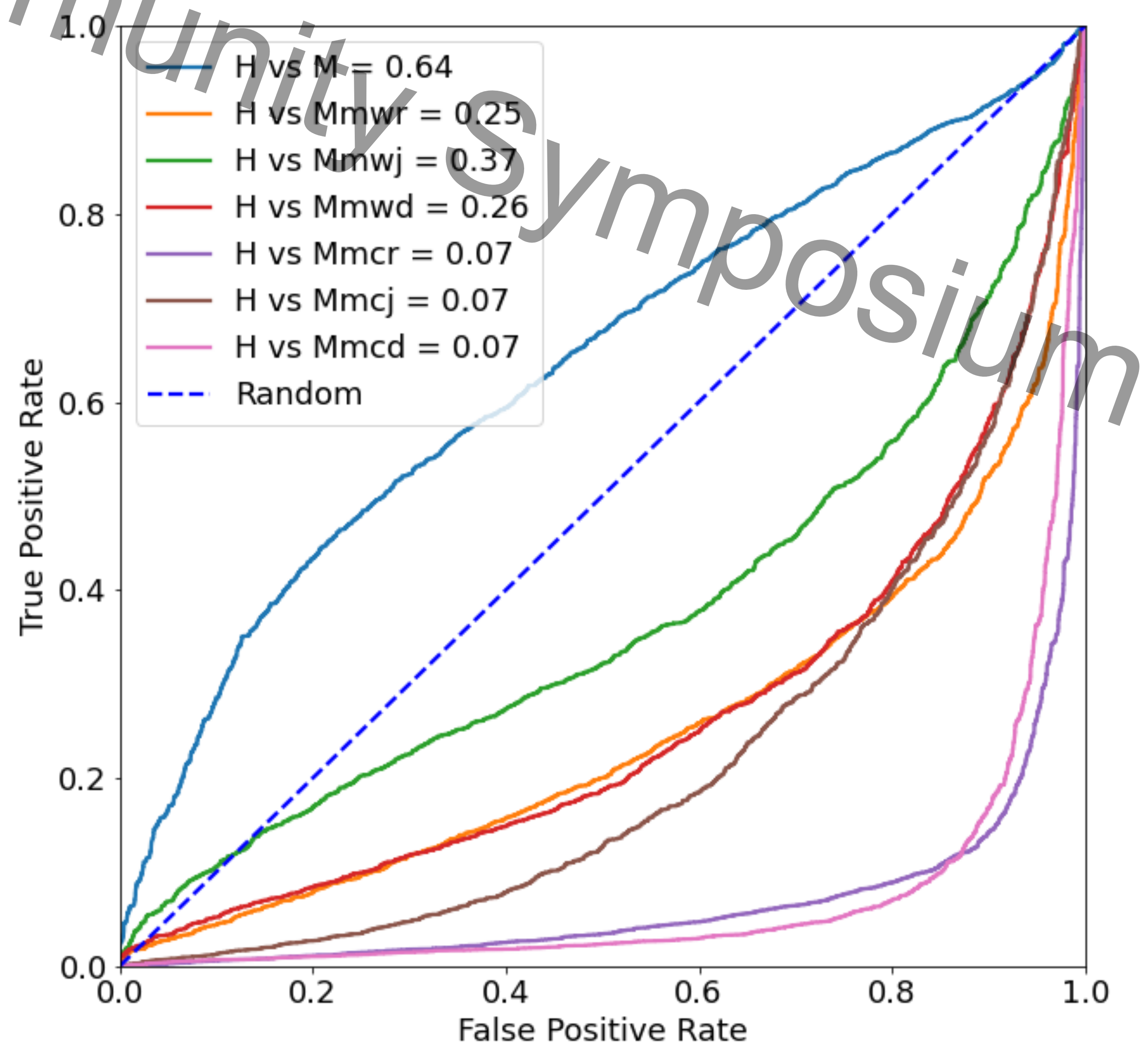
TRUSTWORTHY ISSUES WITH NEURAL NETWORKS

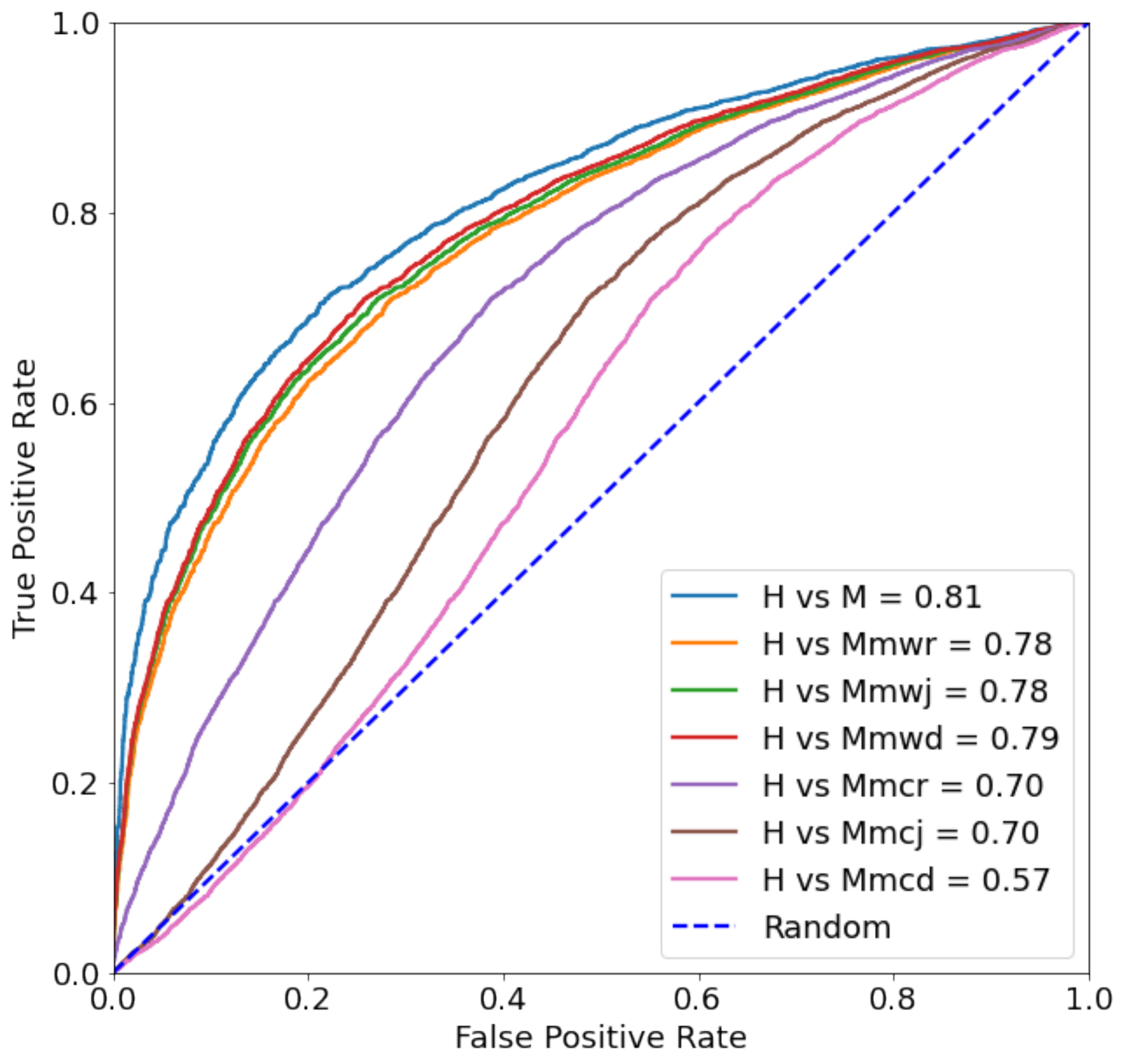
NOT AN EXHAUSTIVE LIST

- **Adversarial attacks:** attempts to trick predictive models into making incorrect predictions or decisions

Mutation Operator	Mutated Sentence
Random misspelling words	**Plz** share and like the video
Random deleting articles	Please share and like ~~**the**~~ video
Random replacing a word with another one	Please **roar** and like the video
Random replacing a word with its synonym	Please **disseminate** and like the video
Random replacing a word with its antonym	Please share and **hate** the video
Random replacing "a" with "α"	Please **shαre** and like the video
Random replacing "e" with "ϵ"	Please share and **likϵ** the **vidϵo**

Input Type	Pre-Trained Model [23]	White-Box Fine-Tuned [23]	Black-Box Fine-Tuned
Human	88.80%	93.65%	92.65 ± 1.04%
Replace Alpha/Epsilon	01.01%	99.92%	99.00 ± 0.98%
Misspelling	00.00%	99.80%	99.49 ± 0.30%
Delete articles	01.60%	25.42%	36.56 ± 5.79%
Synonym replacement	00.00%	99.76%	99.08 ± 0.64%
Random word replacement	07.79%	98.43%	54.44 ± 13.40%
Antonym replacement	09.89%	92.73%	93.74 ± 4.18%



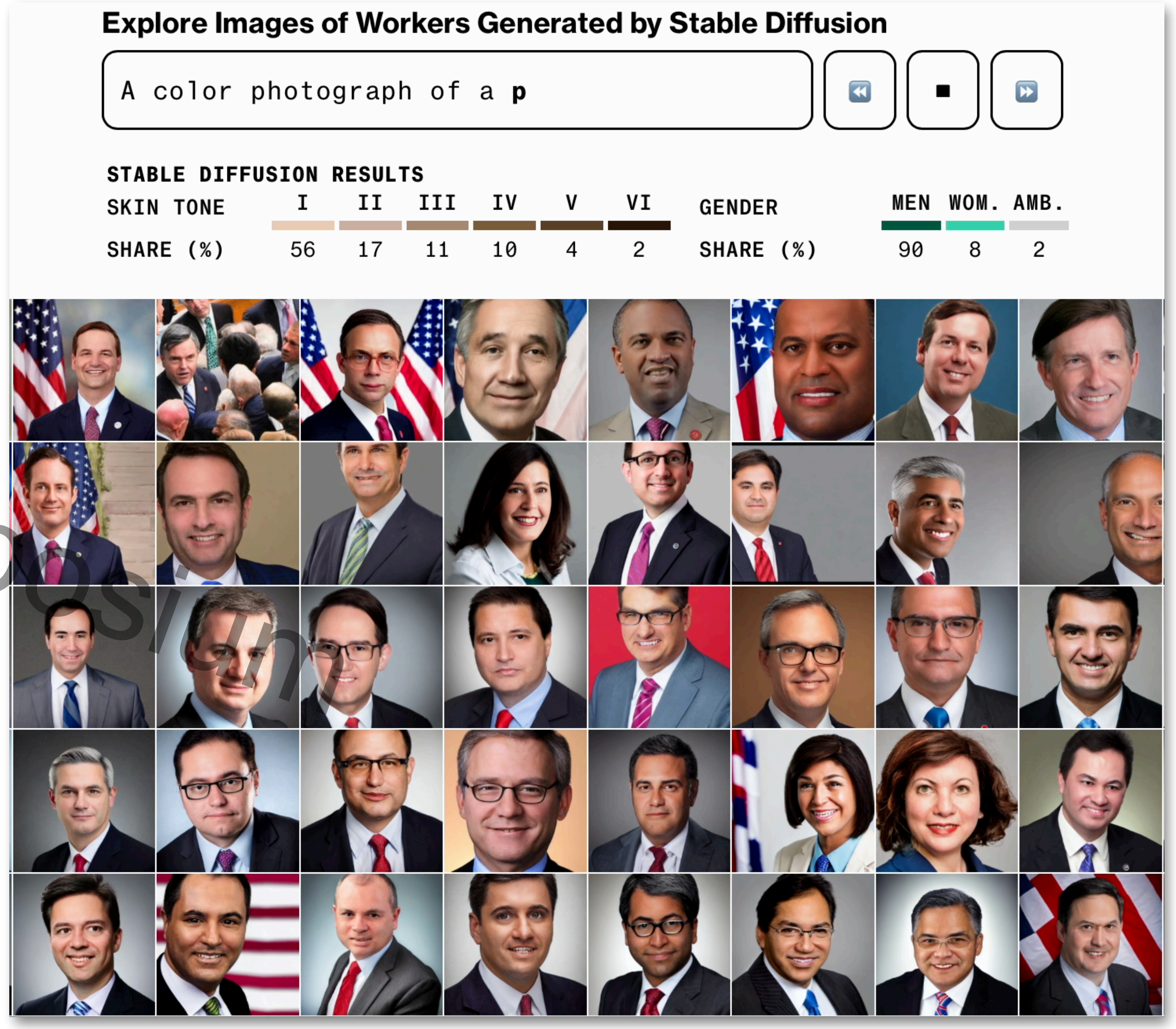


Liang, Gongbo, et al. "Enhancing neural text detector robustness with μ attacking and rr-training." Electronics 12.8 (2023): 1948.
Deanda, Demetrio, et al. "Defending mutation-based adversarial text perturbation: a black-box approach." Cluster Computing 28.3 (2025): 196.

TRUSTWORTHY ISSUES WITH NEURAL NETWORKS

NOT AN EXHAUSTIVE LIST

- **Bias** (e.g., image generative models may associate professions with their stereotypes[1])

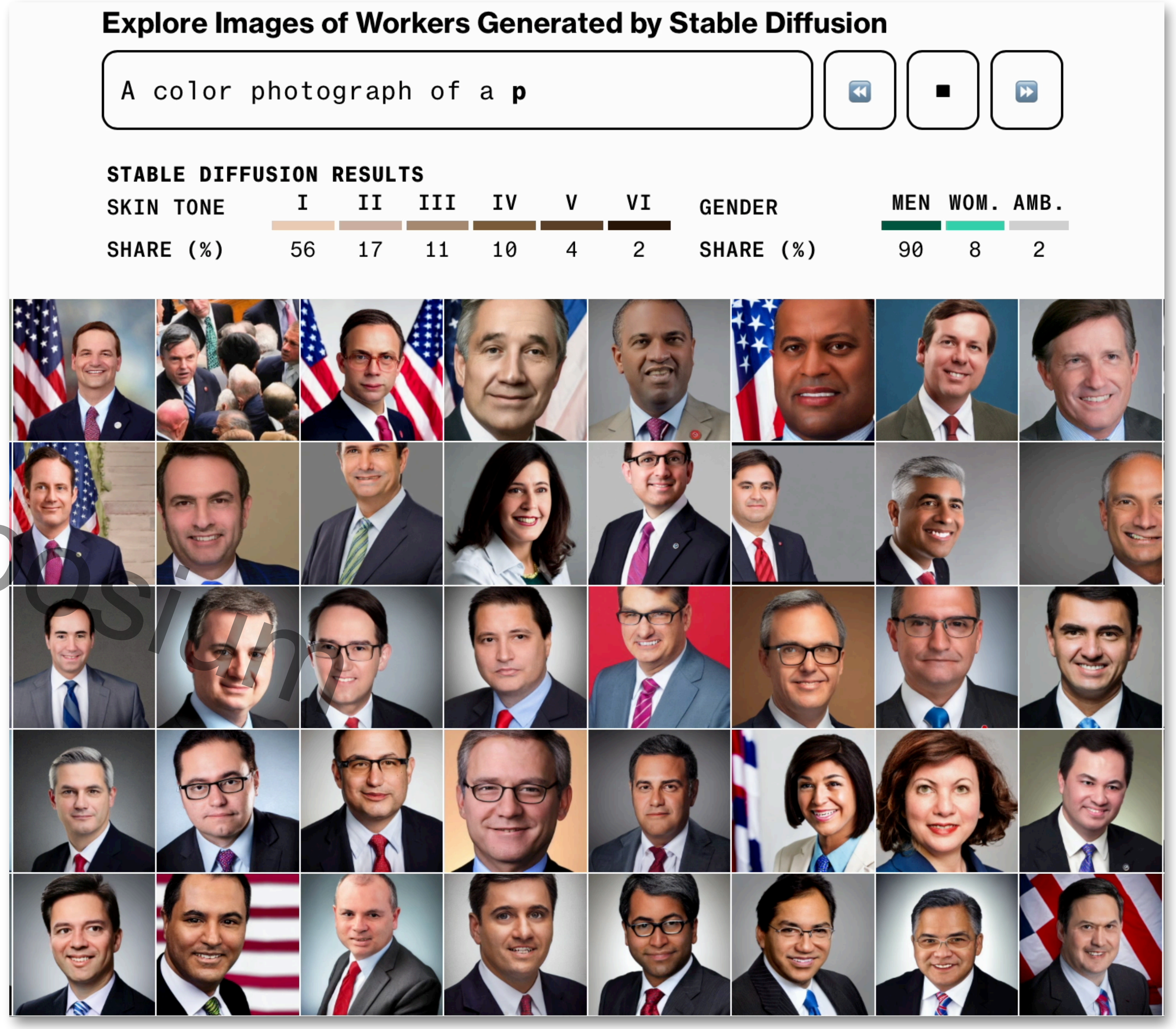


1. https://www.bloomberg.com/graphics/2023-generative-ai-bias/
2. https://interface.media/blog/2024/03/06/big-data-isnt-big-enough-to-train-generative-ai/#:~:text="The development and effectiveness of,billion words—to train ChatGPT.
3. Wang, Xiaoqin, et al. "Inconsistent performance of deep learning models on mammogram classification." Journal of the American College of Radiology 17.6 (2020): 796-803.
4. Xing, Eric, et al. "Neural network decision-making criteria consistency analysis via inputs sensitivity." 26th International Conference on Pattern Recognition (ICPR). IEEE, 2022.

TRUSTWORTHY ISSUES WITH NEURAL NETWORKS

NOT AN EXHAUSTIVE LIST

- **Bias** (e.g., image generative models may associate professions with their stereotypes[1])
- **Data hungry** (e.g., about 300 billion words[2], one billion pages, or 3.5 million books for ChatGPT training)

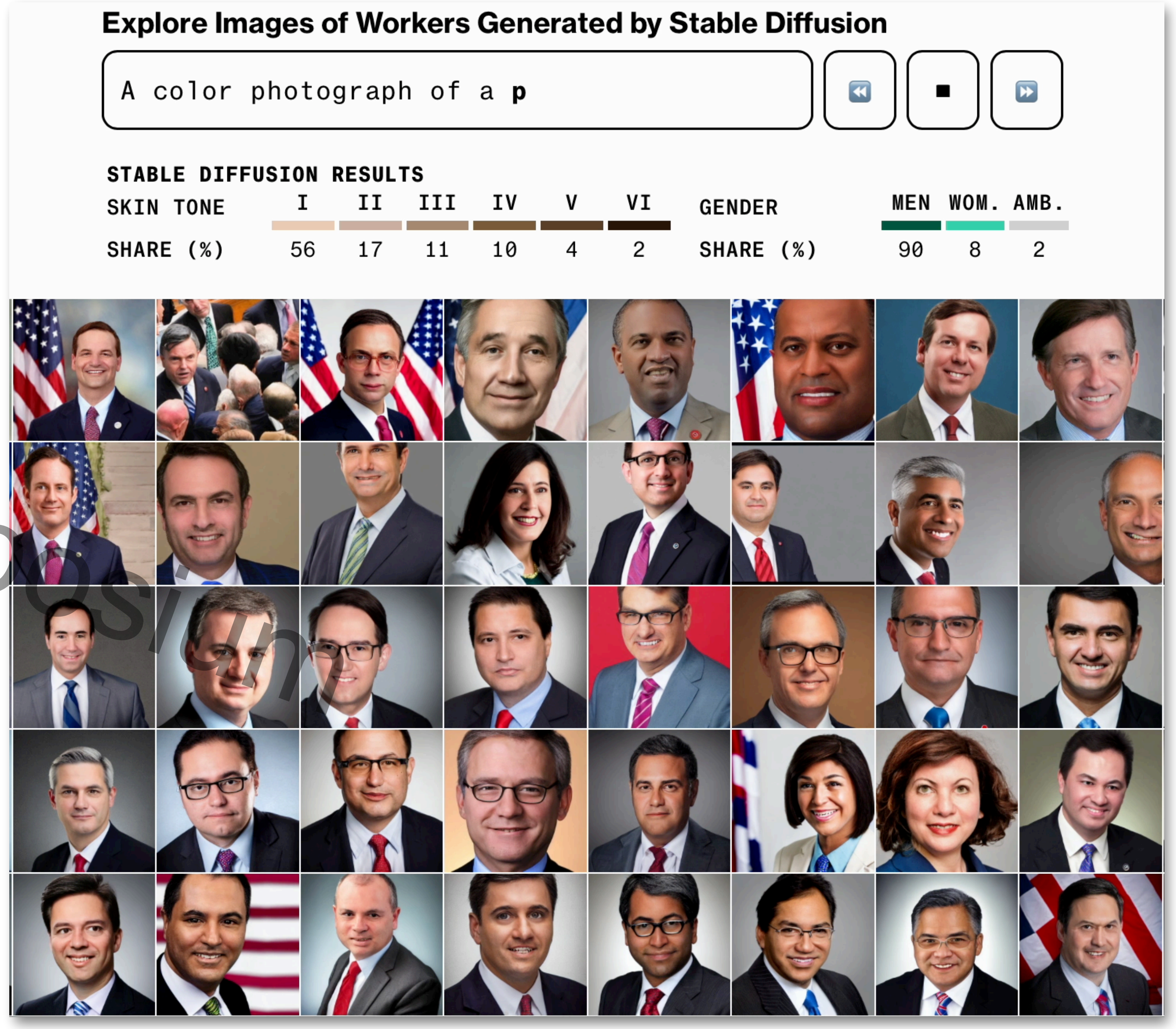


1. https://www.bloomberg.com/graphics/2023-generative-ai-bias/
2. https://interface.media/blog/2024/03/06/big-data-isnt-big-enough-to-train-generative-ai/#:~:text="The development and effectiveness of,billion words—to train ChatGPT.
3. Wang, Xiaoqin, et al. "Inconsistent performance of deep learning models on mammogram classification." Journal of the American College of Radiology 17.6 (2020): 796-803.
4. Xing, Eric, et al. "Neural network decision-making criteria consistency analysis via inputs sensitivity." 26th International Conference on Pattern Recognition (ICPR). IEEE, 2022.

TRUSTWORTHY ISSUES WITH NEURAL NETWORKS

NOT AN EXHAUSTIVE LIST

- **Bias** (e.g., image generative models may associate professions with their stereotypes[1])
- **Data hungry** (e.g., about 300 billion words[2], one billion pages, or 3.5 million books for ChatGPT training)
- **Poor generalization** (e.g., breast cancer model trained for South America does not work well for North America[3])



1. https://www.bloomberg.com/graphics/2023-generative-ai-bias/
2. https://interface.media/blog/2024/03/06/big-data-isnt-big-enough-to-train-generative-ai/#:~:text="The development and effectiveness of,billion words—to train ChatGPT.
3. Wang, Xiaoqin, et al. "Inconsistent performance of deep learning models on mammogram classification." Journal of the American College of Radiology 17.6 (2020): 796-803.
4. Xing, Eric, et al. "Neural network decision-making criteria consistency analysis via inputs sensitivity." 26th International Conference on Pattern Recognition (ICPR). IEEE, 2022.

TRUSTWORTHY ISSUES WITH NEURAL NETWORKS

NOT AN EXHAUSTIVE LIST

- **Bias** (e.g., image generative models may associate professions with their stereotypes[1])
- **Data hungry** (e.g., about 300 billion words[2], one billion pages, or 3.5 million books for ChatGPT training)
- **Poor generalization** (e.g., breast cancer model trained for South America does not work well for North America[3])
- **Low reproducibility** (e.g., it is very hard to train models that are exactly the same[4])

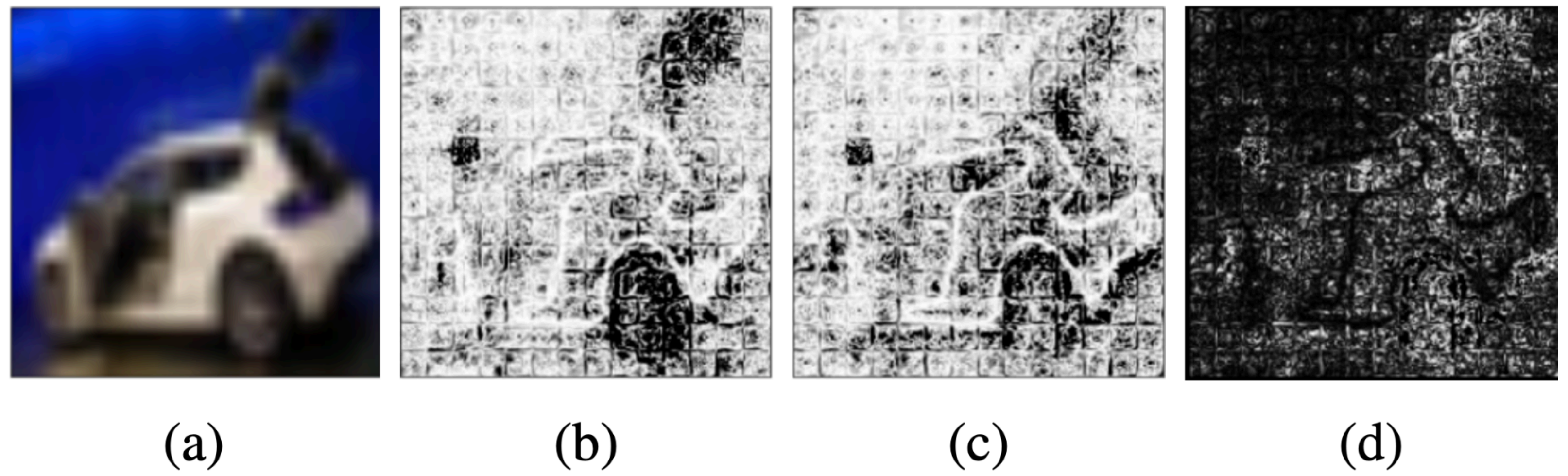
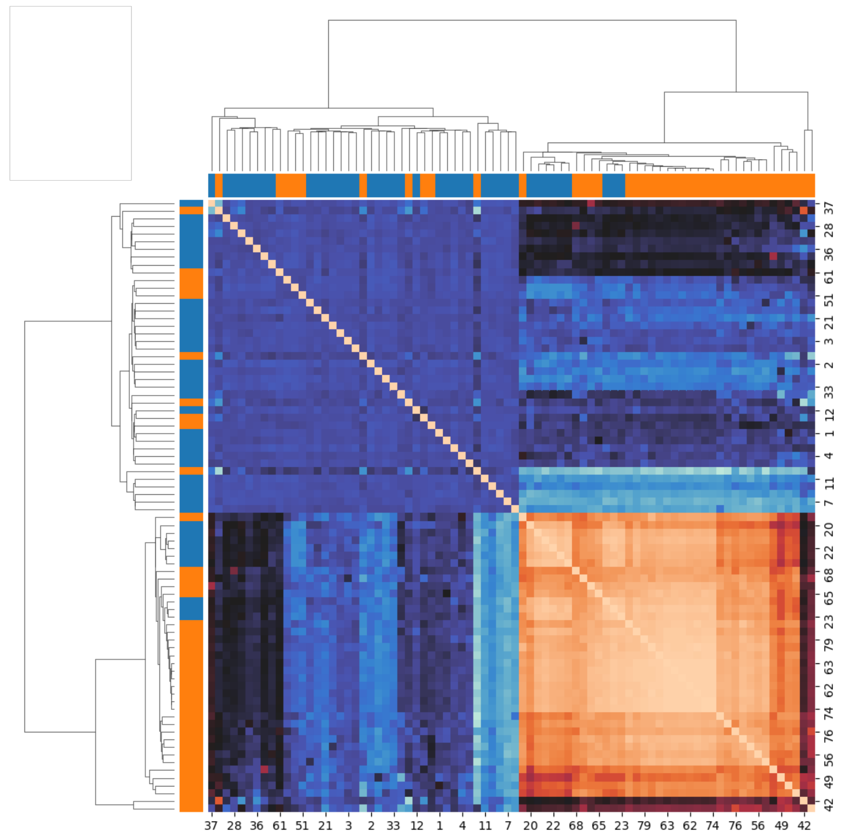


Fig. 1: Feature visualization for two ViT models trained using the same architecture, training data, and hyperparameters. (a) Input image. (b)-(c) Integrated Gradients for Model I and II, show the two models using different features for decision-making making. Darker color indicates more important features. (d) Difference between the features used by the two models.

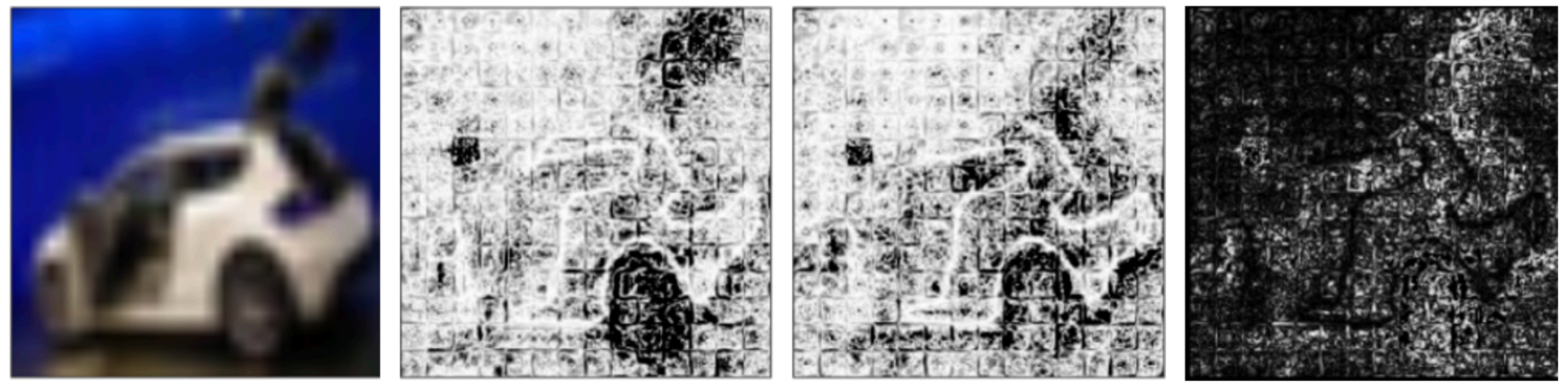


1. https://www.bloomberg.com/graphics/2023-generative-ai-bias/
2. https://interface.media/blog/2024/03/06/big-data-isnt-big-enough-to-train-generative-ai/#:~:text="The development and effectiveness of,billion words—to train ChatGPT.
3. Wang, Xiaoqin, et al. "Inconsistent performance of deep learning models on mammogram classification." Journal of the American College of Radiology 17.6 (2020): 796-803.
4. Xing, Eric, et al. "Neural network decision-making criteria consistency analysis via inputs sensitivity." 26th International Conference on Pattern Recognition (ICPR). IEEE, 2022.

TRUSTWORTHY ISSUES WITH NEURAL NETWORKS

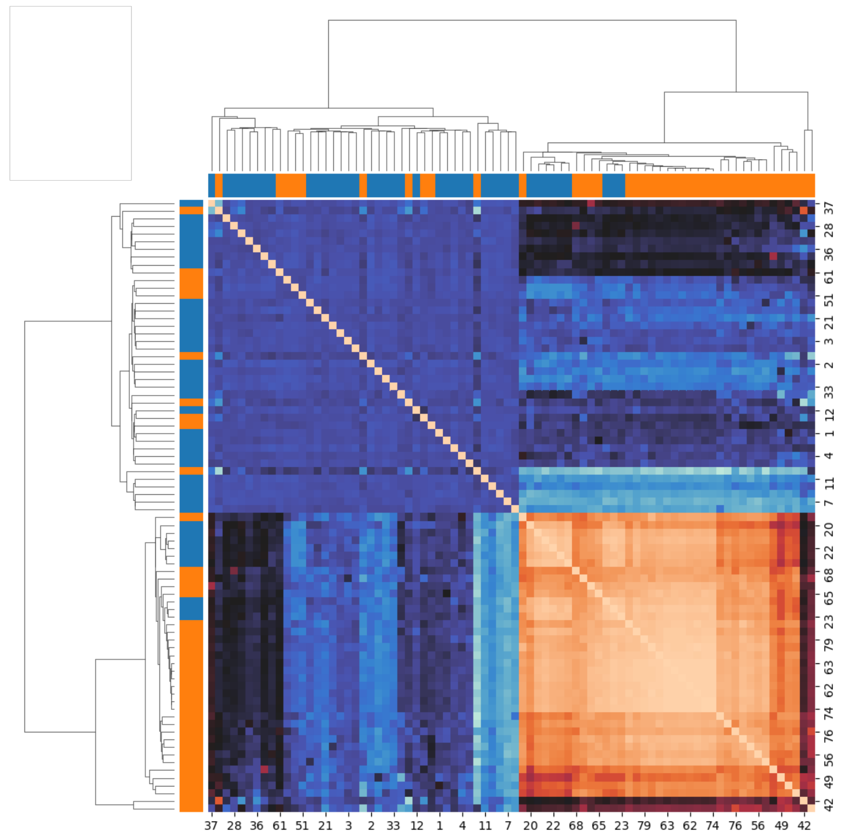
NOT AN EXHAUSTIVE LIST

- **Bias** (e.g., image generative models may associate professions with their stereotypes[1])
- **Data hungry** (e.g., about 300 billion words[2], one billion pages, or 3.5 million books for ChatGPT training)
- **Poor generalization** (e.g., breast cancer model trained for South America does not work well for North America[3])
- **Low reproducibility** (e.g., it is very hard to train models that are exactly the same[4])
- **Poor calibration** (e.g., Death of Elaine Herzberg)



(a) (b) (c) (d)

Fig. 1: Feature visualization for two ViT models trained using the same architecture, training data, and hyperparameters. (a) Input image. (b)-(c) Integrated Gradients for Model I and II, show the two models using different features for decision-making making. Darker color indicates more important features. (d) Difference between the features used by the two models.



1. https://www.bloomberg.com/graphics/2023-generative-ai-bias/
2. https://interface.media/blog/2024/03/06/big-data-isnt-big-enough-to-train-generative-ai/#:~:text="The development and effectiveness of,billion words—to train ChatGPT.
3. Wang, Xiaoqin, et al. "Inconsistent performance of deep learning models on mammogram classification." Journal of the American College of Radiology 17.6 (2020): 796-803.
4. Xing, Eric, et al. "Neural network decision-making criteria consistency analysis via inputs sensitivity." 26th International Conference on Pattern Recognition (ICPR). IEEE, 2022.

TRUSTWORTHY ISSUES WITH NEURAL NETWORKS

POTENTIAL TECHNICAL REASON

- **Bias** (e.g., image generative models may associate professions with their stereotypes)
- **Data hungry** (e.g., about 300 billion words, one billion pages, or 3.5 million books for ChatGPT training)
- **Poor generalization** (e.g., breast cancer model trained for South America does not work well for North America)
- **Low reproducibility** (e.g., it is very hard to train models that are exactly the same)
- **Poor calibration** (e.g., Death of Elaine Herzberg)

TRAINING DATA / DATA DISTRIBUTION & LIMITATION OF EXISTING NN MODELS

LEARNING IS ONLY GUIDED BY THE FINAL RESULT

POOR UNCERTAINTY ESTIMATION ABILITY

GONGBO "TONY" LIANG, PHD (GLIANG@TAMUSA.EDU)

ADDRESSING TRUST AND SAFETY CHALLENGES IN NEURAL NETWORK-POWERED MODERN AI:

A call for broader awareness and action

CONCLUSION

Research: The rapid adoption of neural network-based solutions in our daily lives necessitates increased attention to the vulnerabilities of these networks within the cybersecurity domain.

Education: Incorporating trustworthy AI into college curricula could be beneficial, such as general education courses to raise awareness and upper-level courses designed for computing majors.

Enhanced model uncertainty estimation has the potential to significantly improve the trustworthiness of neural networks.