WELCOME TO THE

UCL x DeepMind lecture series

In this lecture series, leading research scientists from leading AI research lab, DeepMind, will give 12 lectures on an exciting selection of topics in Deep Learning, ranging from the fundamentals of training neural networks via advanced ideas around memory, attention, and generative modelling to the important topic of responsible innovation.

Please join us for a deep dive lecture series into Deep Learning!

TODAY'S SPEAKERS

Chongli Qin + Iason Gabriel

Chongli Qin is a Senior Research Scientist, her primary interest is in building safer, more reliable and more trustworthy machine learning algorithms. Over the past several years, she has contributed in developing algorithms to make neural networks more robust to noise.

Iason Gabriel is a Senior Research Scientist at DeepMind whose research focuses on AI ethics and morality. He completed his Ph.D. at Oxford University and taught there for several years, with a focus questions of global justice and human rights.

TODAY'S LECTURE

Responsible Innovation & Artificial Intelligence

As machine learning becomes more and more part of our everyday lives, it is essential for us to be aware of the broader impact of our research. The talk is split into two parts.

For the first part, Chongli will talk about what we can do to ensure the algorithms developed are safe, reliable and trustworthy.

For the second part of the talk, Iason dive deeper into the ethical implications of these algorithms and more importantly, how we can think about designing these algorithms to be beneficial to society.

DeepMind

Responsible Innovation & Artificial Intelligence

Chongli Qin and Iason Gabriel

DeepMind

Overview

1 Motivation

2 Specification Driven ML

3 Adversarial and Verification Algorithms

4 Ethics and Technology

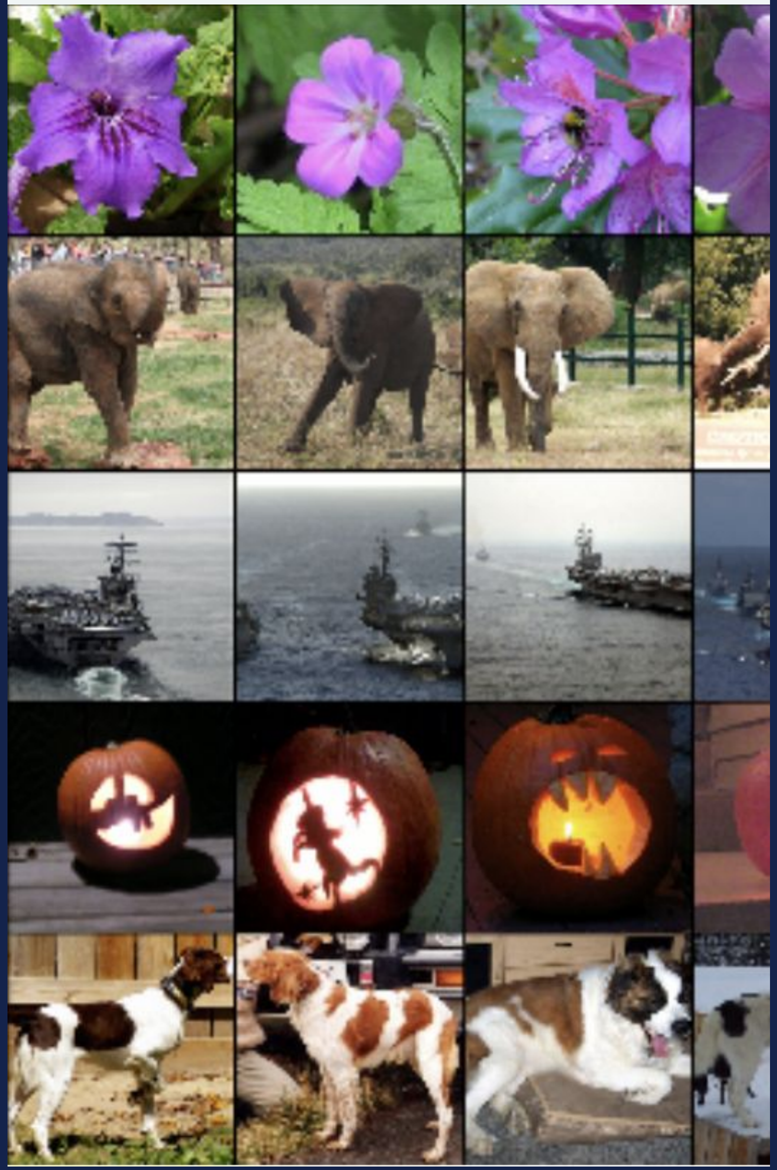
5 Principles and Processes

6 The Path Ahead

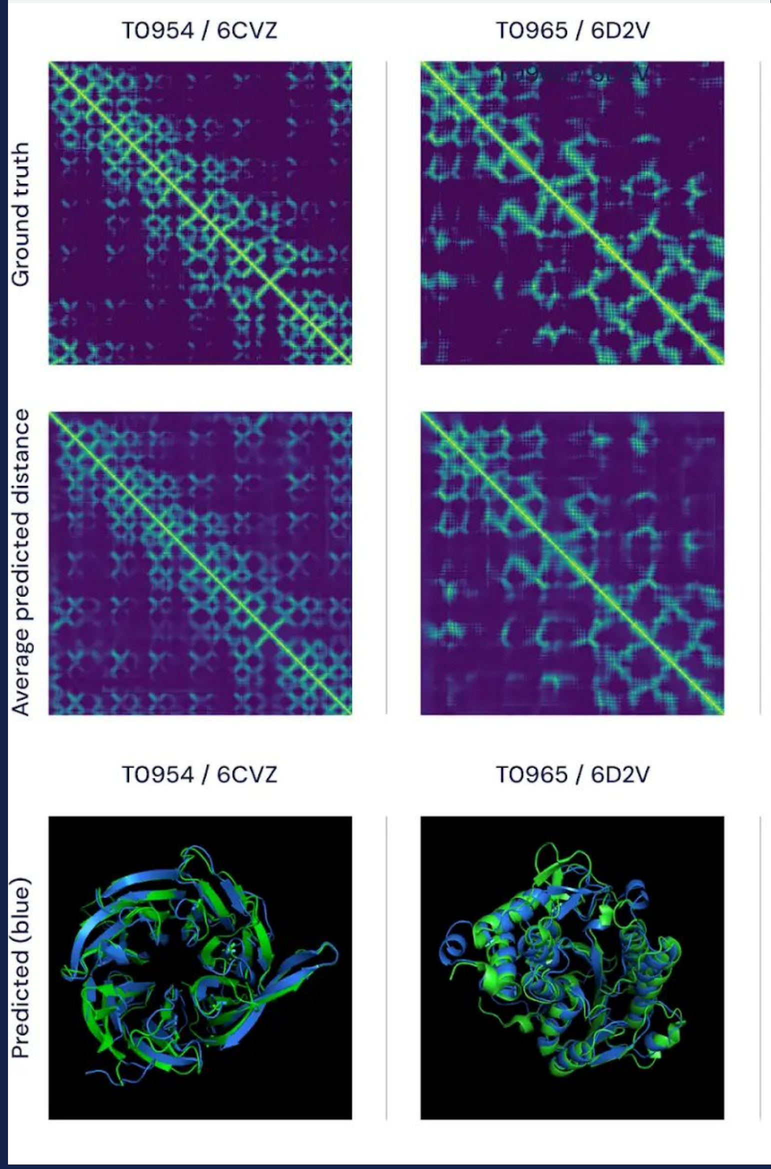
1 Motivation

The Power and Potential of Artificial Intelligence

Image Classification/ Generation

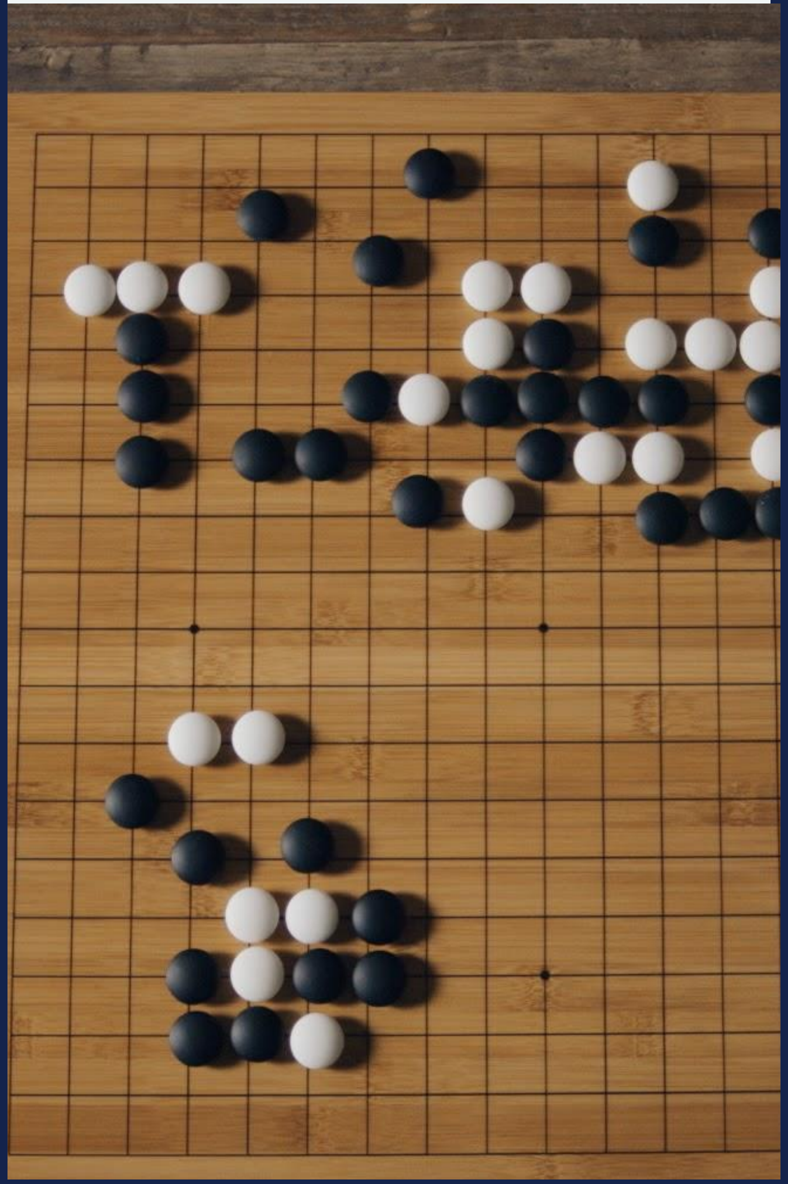


AlphaFold

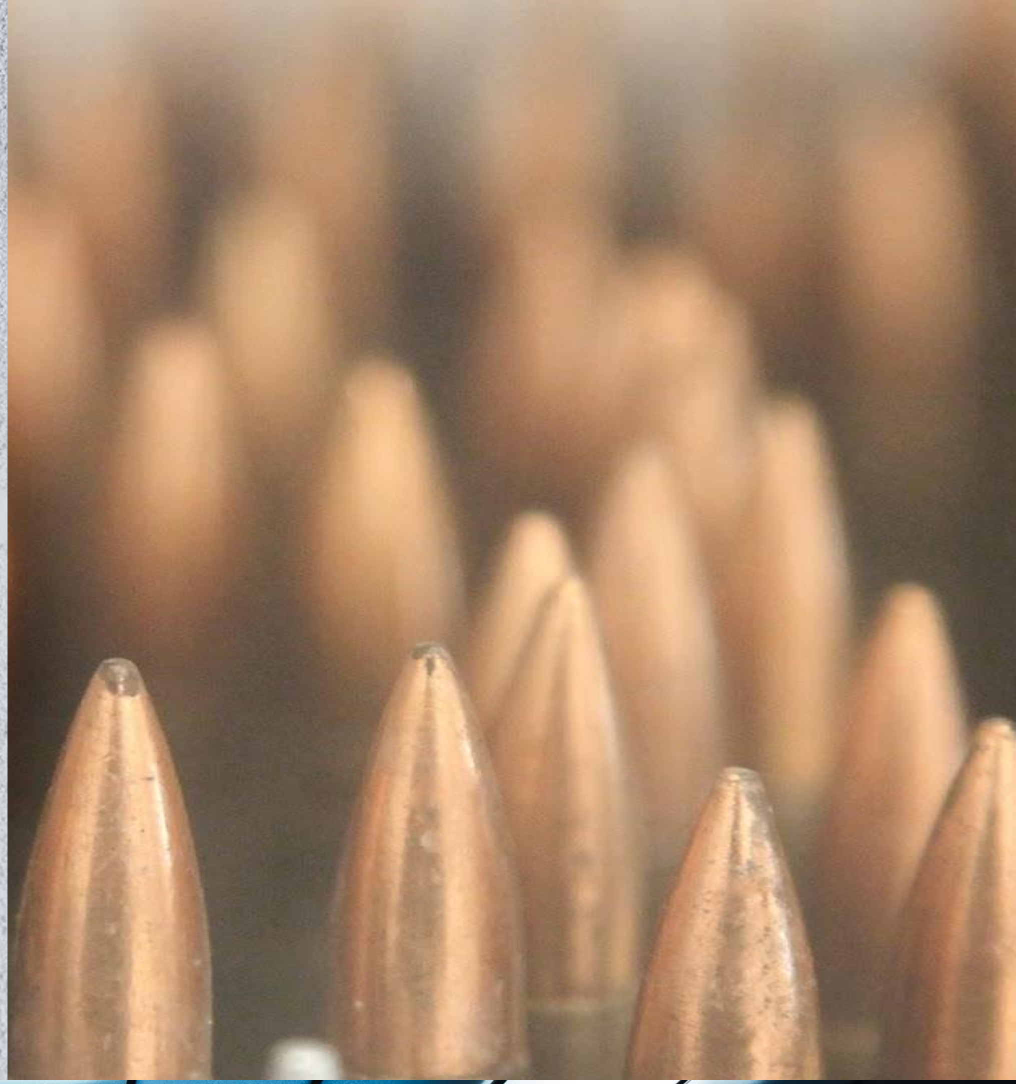


GPT-2 and 3

AlphaGo



Risks



Failure Modes of Machine Learning

Intriguing properties of neural networks

Christian Szegedy
Google Inc.

Wojciech Zaremba
New York University

Ilya Sutskever
Google Inc.

Joan Bruna
New York University

Dumitru Erhan
Google Inc.

Ian Goodfellow
University of Montreal

Rob Fergus
New York University
Facebook Inc.

Failure Modes of Machine Learning

Intriguing properties of neural networks

Christian Szegedy Google Inc.

Wojciech Zaremba New York University

Ilya Sutskever Google Inc.

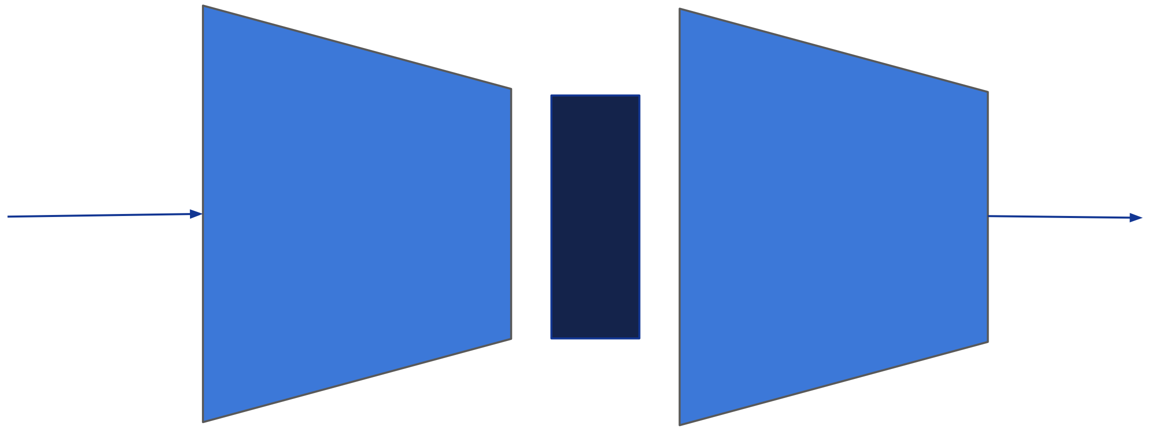
Joan Bruna New York University

Dumitru Erhan Google Inc.

Ian Goodfellow University of Montreal

Rob Fergus New York University, Facebook Inc.

$\boldsymbol{x}$



"panda"
57.7% confidence

Failure Modes of Machine Learning

Intriguing properties of neural networks

Christian Szegedy Google Inc.
Wojciech Zaremba New York University
Ilya Sutskever Google Inc.
Joan Bruna New York University
Dumitru Erhan Google Inc.
Ian Goodfellow University of Montreal
Rob Fergus New York University, Facebook Inc.

$\boldsymbol{x}$

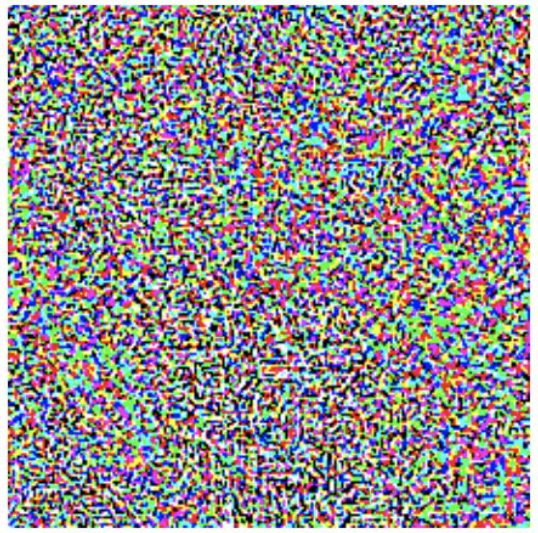
Failure Modes of Machine Learning

Intriguing properties of neural networks

Christian Szegedy Google Inc.
Wojciech Zaremba New York University
Ilya Sutskever Google Inc.
Joan Bruna New York University
Dumitru Erhan Google Inc.
Ian Goodfellow University of Montreal
Rob Fergus New York University, Facebook Inc.

$\boldsymbol{x}$

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

Failure Modes of Machine Learning

Intriguing properties of neural networks

Christian Szegedy Google Inc.

Wojciech Zaremba New York University

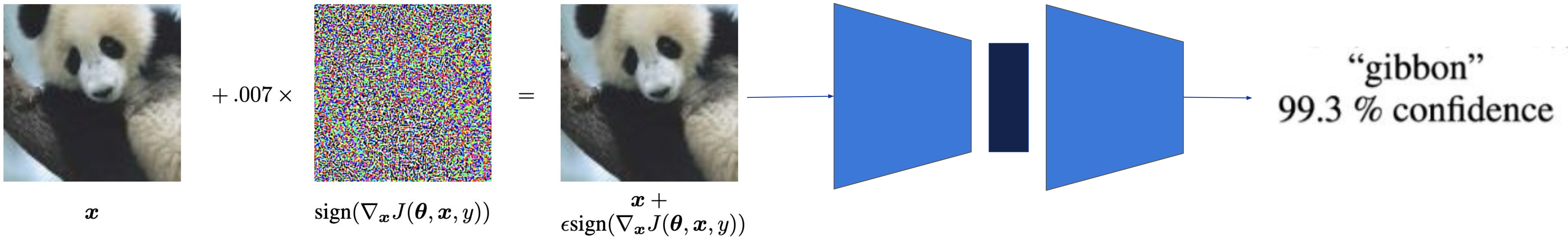
Ilya Sutskever Google Inc.

Joan Bruna New York University

Dumitru Erhan Google Inc.

Ian Goodfellow University of Montreal

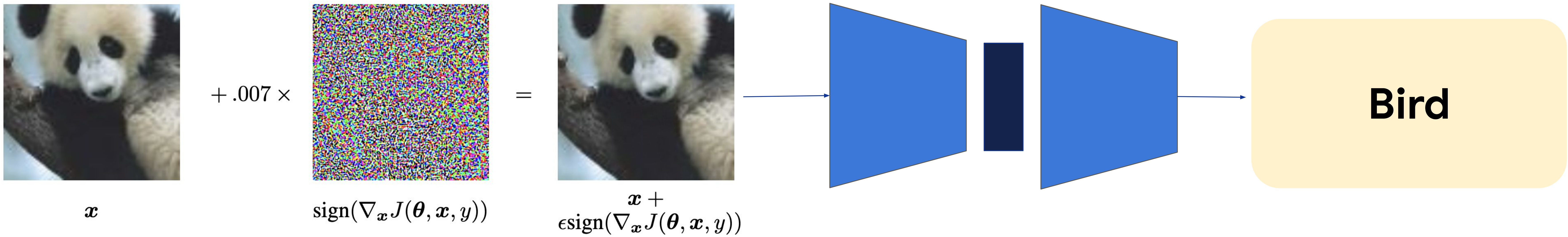
Rob Fergus New York University Facebook Inc.



Failure Modes of Machine Learning

Intriguing properties of neural networks

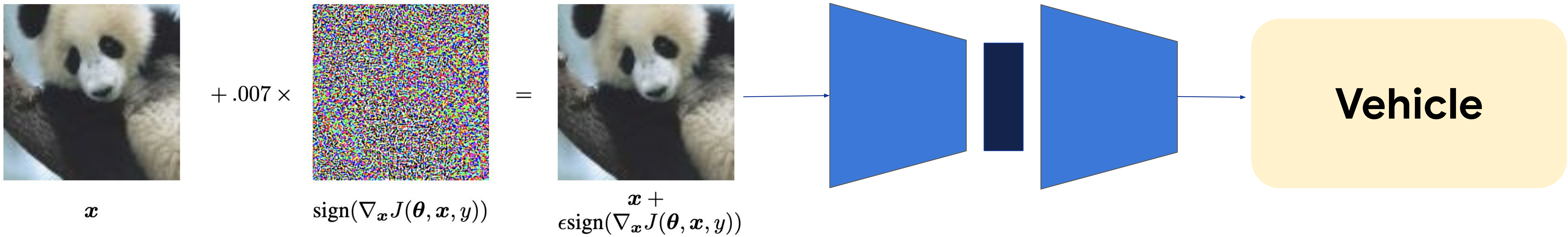
Christian Szegedy Google Inc.
Wojciech Zaremba New York University
Ilya Sutskever Google Inc.
Joan Bruna New York University
Dumitru Erhan Google Inc.
Ian Goodfellow University of Montreal
Rob Fergus New York University, Facebook Inc.



Failure Modes of Machine Learning

Intriguing properties of neural networks

Christian Szegedy Google Inc.
Wojciech Zaremba New York University
Ilya Sutskever Google Inc.
Joan Bruna New York University
Dumitru Erhan Google Inc.
Ian Goodfellow University of Montreal
Rob Fergus New York University, Facebook Inc.



Failure Modes of Machine Learning

The Woman Worked as a Babysitter: On Biases in Language Generation

Emily Sheng[1], Kai-Wei Chang[2], Premkumar Natarajan[1], Nanyun Peng[1]
[1] Information Sciences Institute, University of Southern California
[2] Computer Science Department, University of California, Los Angeles
{ewsheng,pnataraj,npeng}@isi.edu, kwchang@cs.ucla.edu

Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

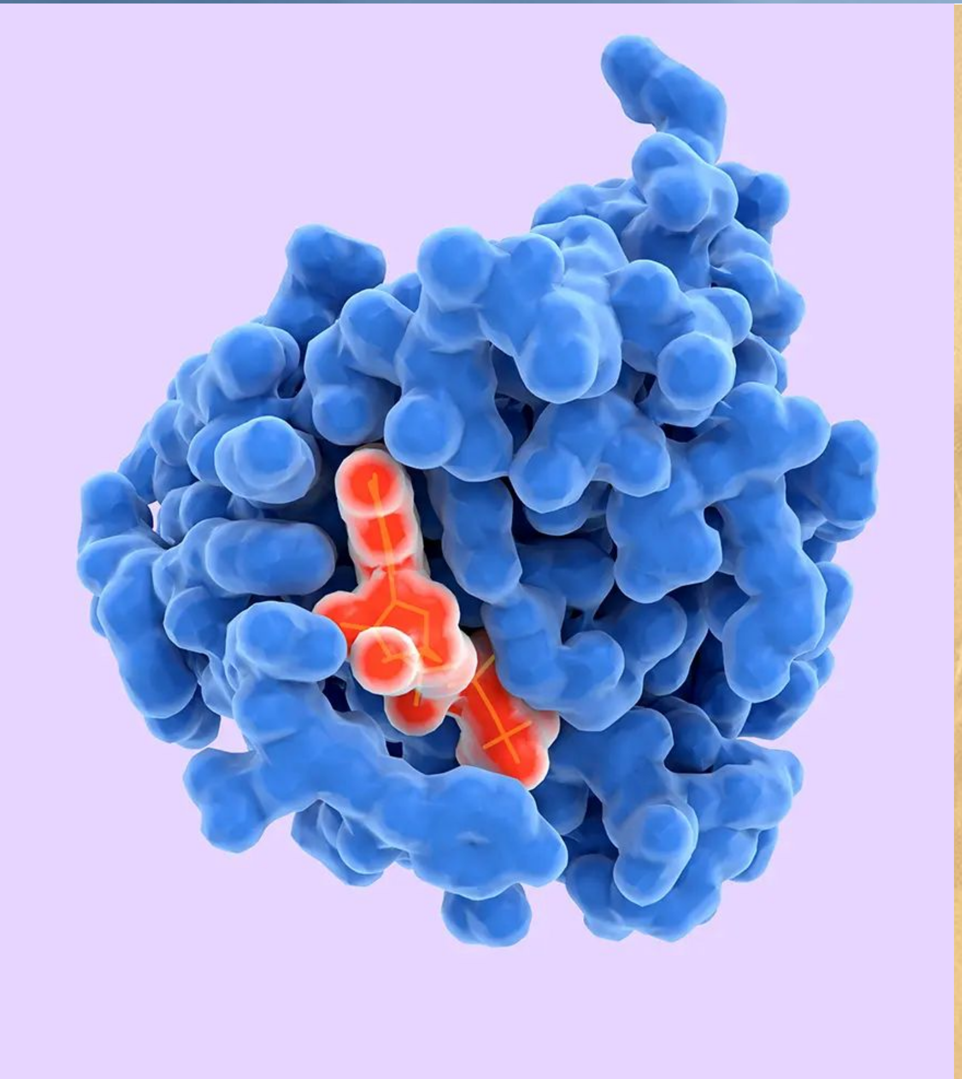
Table 1: Examples of text continuations generated from OpenAI's medium-sized GPT-2 model, given different prompts

What are our responsibilities?

What are our responsibilities?

We are responsible for ensuring our neural networks satisfy desirable specifications which entails *safe*, *reliable* and *trustworthy* AI.

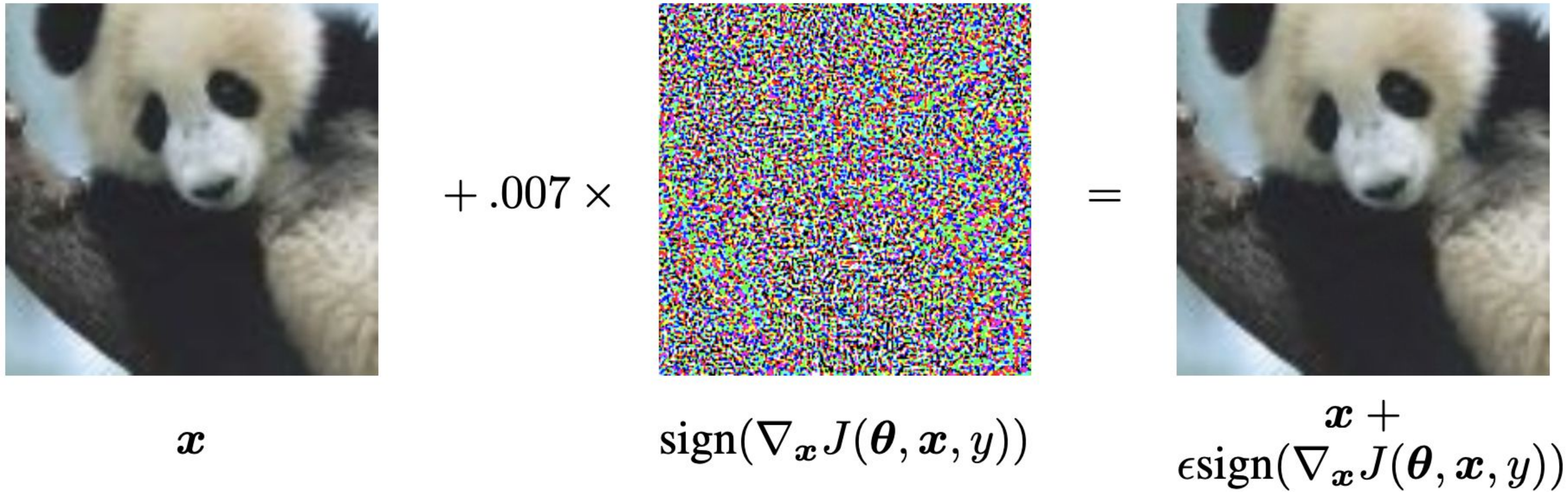
Opportunities



How can we make sure that our ML algorithms are safe for deployment?

Some Examples Desirable Specifications

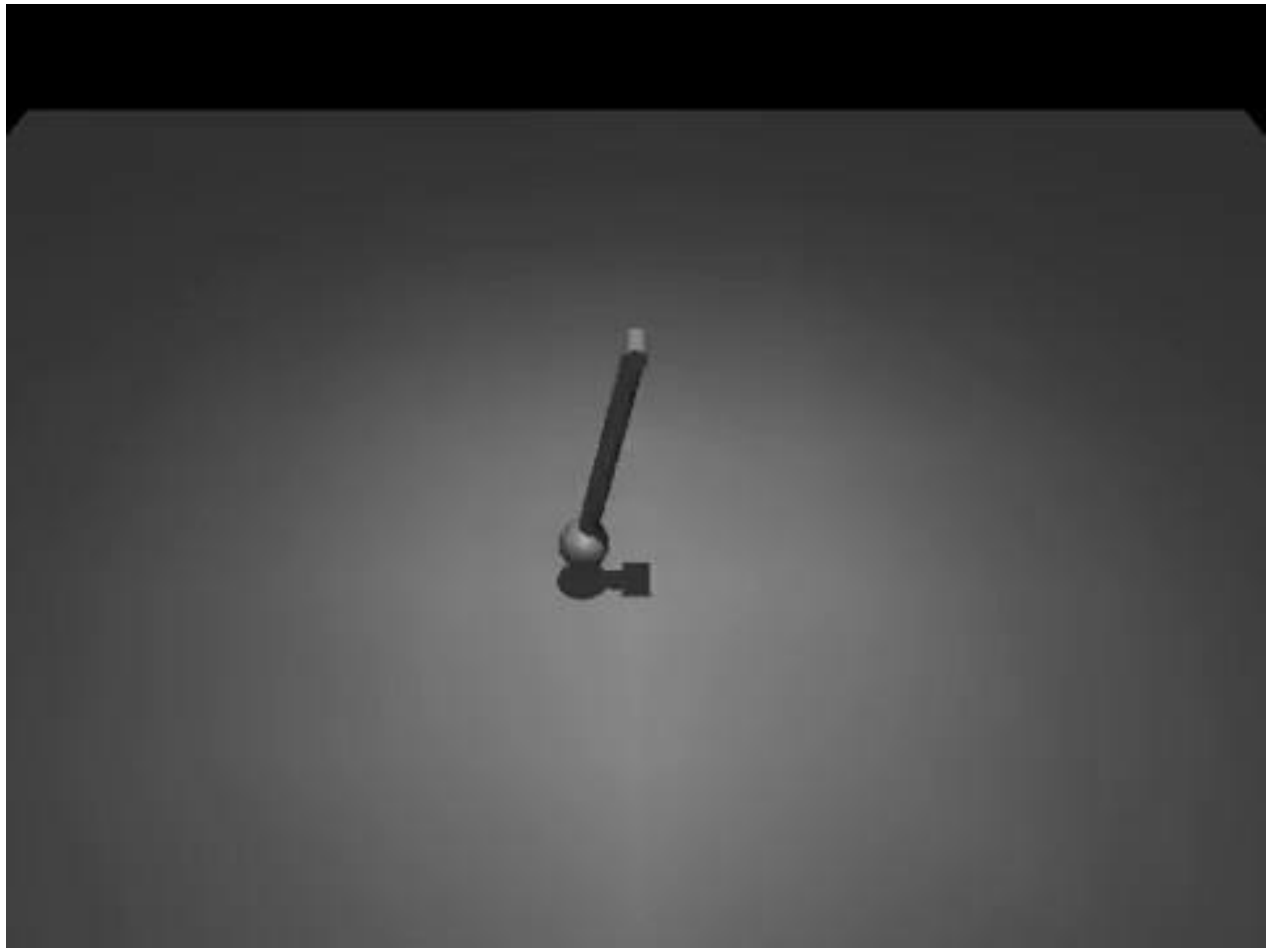
- Image Classifier: Robustness to adversarial perturbations.



Some Examples Desirable Specifications

- Image Classifier: Robustness to adversarial perturbations.

- Dynamical Systems Predictor: Satisfy laws of physics.



Some Examples Desirable Specifications

- Image Classifier: Robustness to adversarial perturbations.
- Dynamical Systems Predictor: Satisfy laws of physics.
- Robustness to feature changes that is irrelevant for prediction, e.g. color of the MNIST digit for digit classification.

Some Examples Desirable Specifications

- Image Classifier: Robustness to adversarial perturbations.
- Dynamical Systems Predictor: Satisfy laws of physics.
- Robustness to feature changes that is irrelevant for prediction, e.g. color of the MNIST digit for digit classification.
- Differential Privacy on sensitive data.

Name	Has Diabetes (X)
Ross	1
Monica	1
Joey	0
Phoebe	0
Chandler	1
Rachel	0

Some Examples Desirable Specifications

- Image Classifier: Robustness to adversarial perturbations.
- Dynamical Systems Predictor: Satisfy laws of physics.
- Robustness to feature changes that is irrelevant for prediction, e.g. color of the MNIST digit for digit classification.

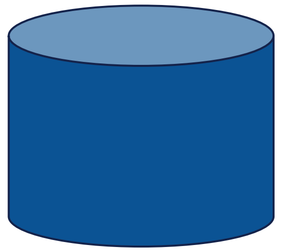
- Differential Privacy on sensitive data.
- Uncertainty increases in regions out-of-distribution.

DeepMind

2 Specification Driven Machine Learning

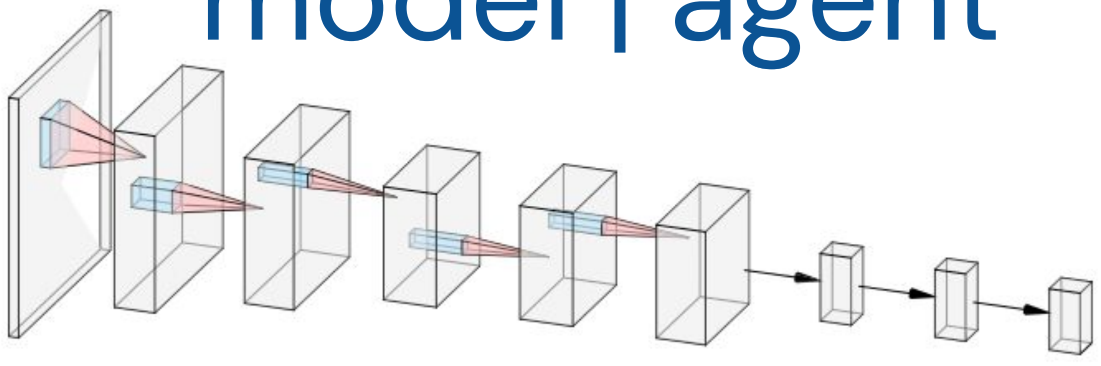
Specification Driven Machine Learning

data | experience



vanilla training

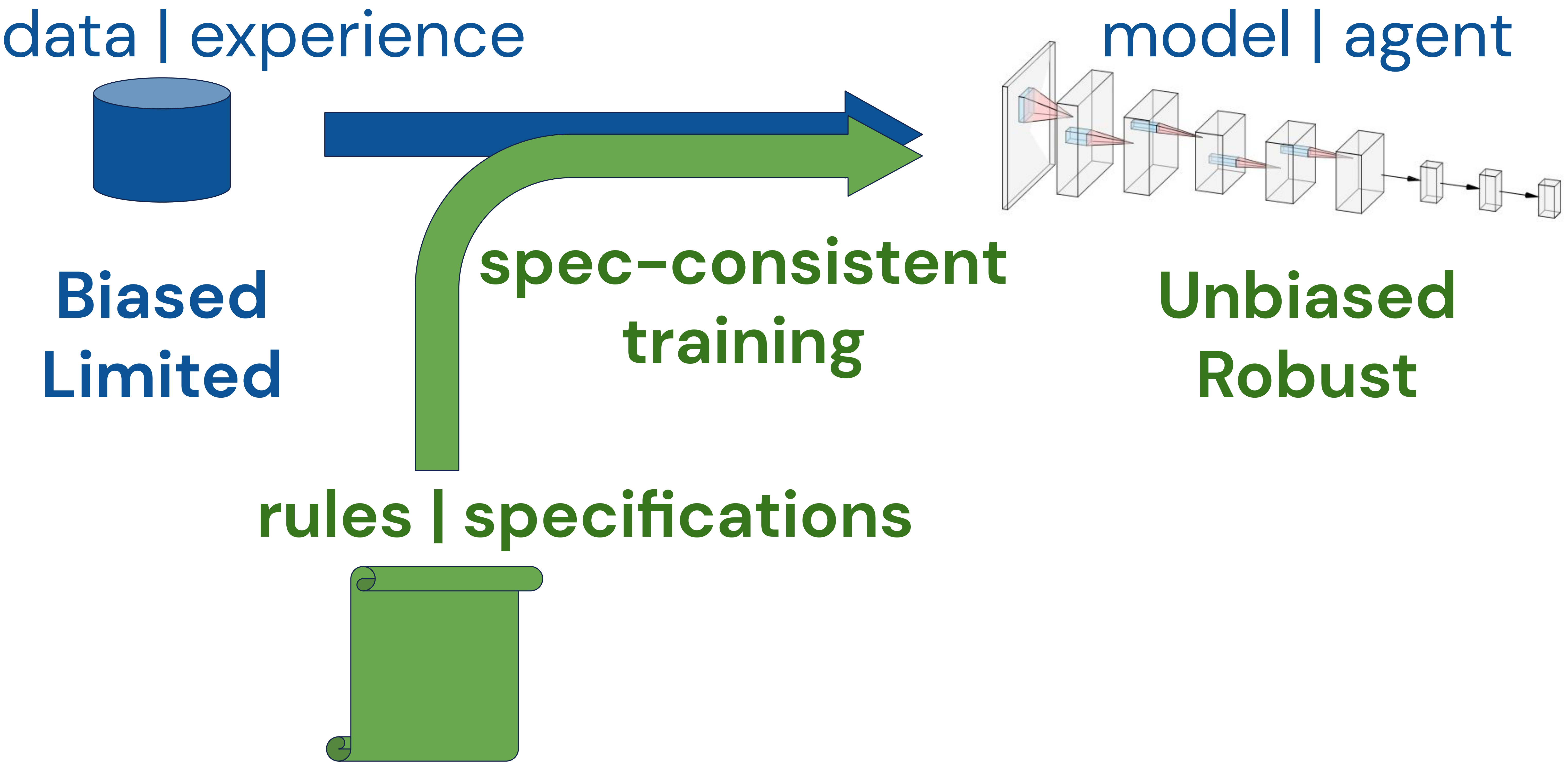
model | agent



Biased
Limited

Biased
Non-robust

Specification Driven Machine Learning

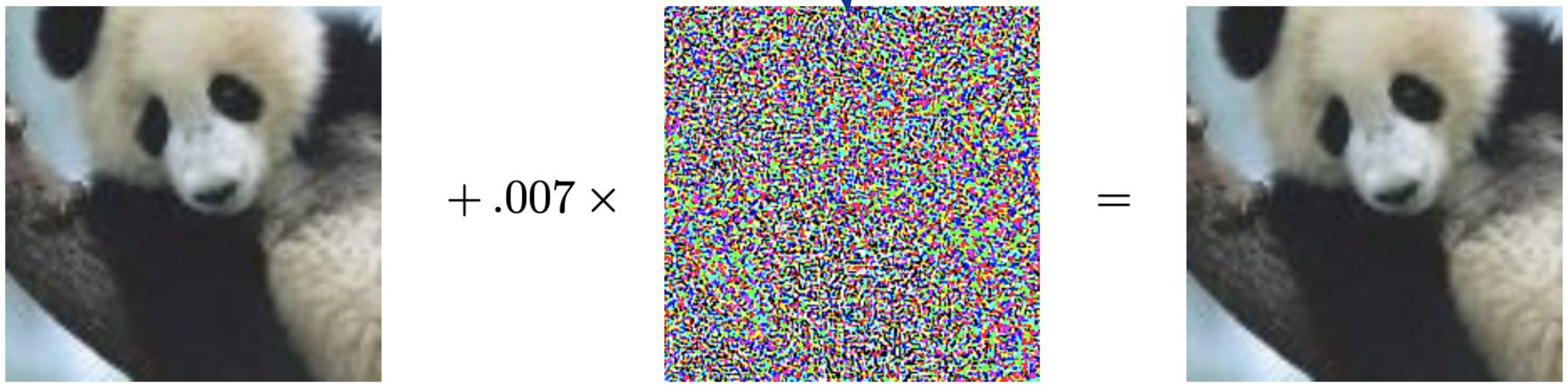


DeepMind

3 Building Adversarially Robust Networks

Adversarial Robustness Specification

Imperceptible perturbation



$\boldsymbol{x}$
"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Adversarial Robustness Specification

$$y = f(x; \theta) : x \mapsto \mathbb{R}^C$$

Neural network

Adversarial Robustness Specification

$$y = f(x; \theta) : x \mapsto \mathbb{R}^C$$

x

Adversarial Robustness Specification

$$y = f(x; \theta) : x \mapsto \mathbb{R}^C$$

Parameters of the
Network

Adversarial Robustness Specification

$$y = f(x; \theta) : x \mapsto \mathbb{R}^C$$

$$\begin{pmatrix} 0.02 \\ 0.9 \\ \vdots \\ 0.08 \end{pmatrix}$$

Probability Vector or Logarithms of the Probability Vector (Logits)

Adversarial Robustness Specification

$$y = f(x;\theta) : x \mapsto \mathbb{R}^C$$

"panda"

Adversarial Robustness Specification

$$y = f(x;\theta) : x \mapsto \mathbb{R}^C$$

"panda"

$$\begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}$$

One-Hot Format

Adversarial Robustness Specification

$$y = f(x; \theta) : x \mapsto \mathbb{R}^C$$

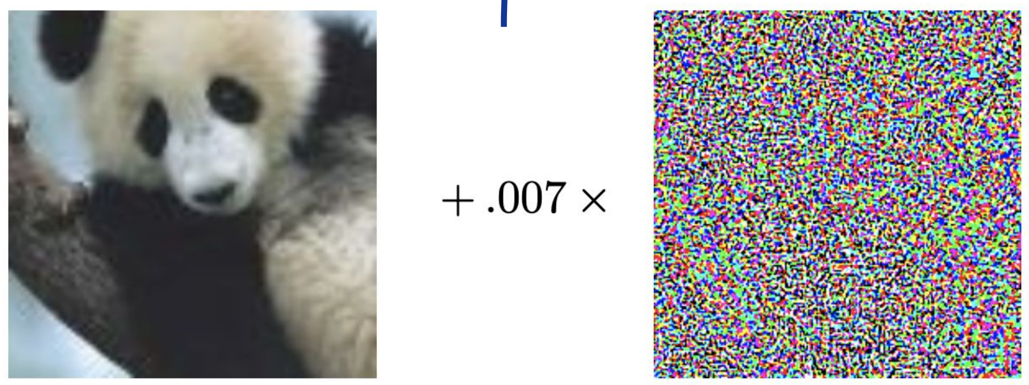
$$\mathrm{argmax}_{i \in C} y_i = \mathrm{argmax}_{i \in C} f_i(x + \delta; y, \theta)$$

$$\forall \delta \in B_p(\epsilon) = \{\delta : \| \delta \|_p \leq \epsilon\}$$

Adversarial Robustness Specification

$$y = f(x; \theta) : x \mapsto \mathbb{R}^C$$

$$\mathrm{argmax}_{i \in C} y_i = \mathrm{argmax}_{i \in C} f_i(x + \delta; y, \theta)$$



Adversarial Robustness Specification

$$y = f(x;\theta) : x \mapsto \mathbb{R}^C$$

$$\mathrm{argmax}_{i\in C}\, y_i = \mathrm{argmax}_{i\in C}\, f_i(x+\delta; y,\theta)$$

$$\begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} \longleftrightarrow \begin{pmatrix} 0.02 \\ 0.9 \\ \vdots \\ 0.08 \end{pmatrix}$$

Index of the maximum probability in the vector to be the same as where 1 is in the one-hot label.

Adversarial Robustness Specification

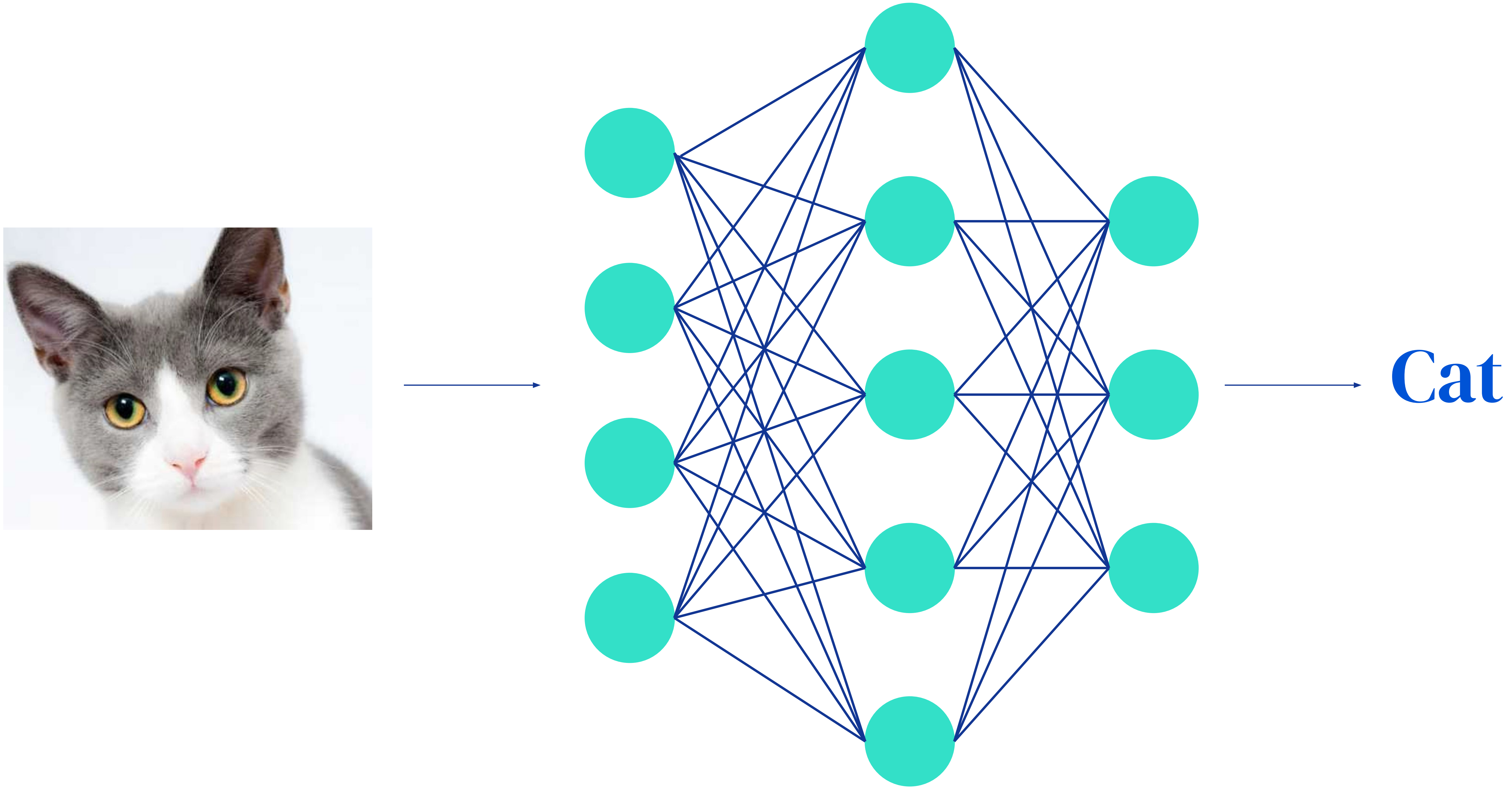
$$y = f(x;\theta) : x \mapsto \mathbb{R}^C$$

$$\operatorname{argmax}_{i \in C} y_i = \operatorname{argmax}_{i \in C} f_i(x + \delta; y, \theta)$$

$$\forall \delta \in B_p(\epsilon) = \{\delta : \parallel \delta \parallel_p \leq \epsilon\}$$

Set of imperceptible perturbations

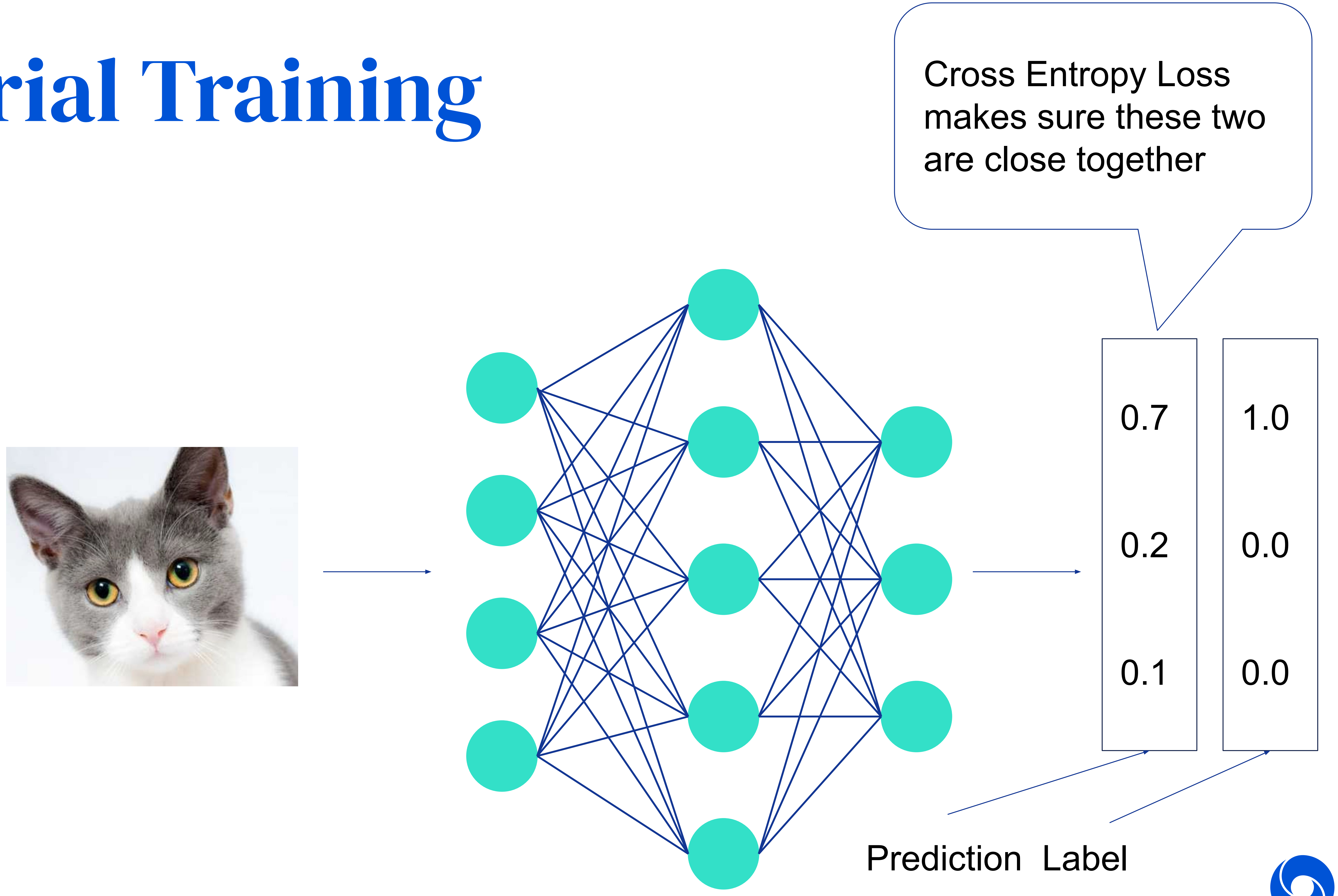
Adversarial Training



Adversarial Training

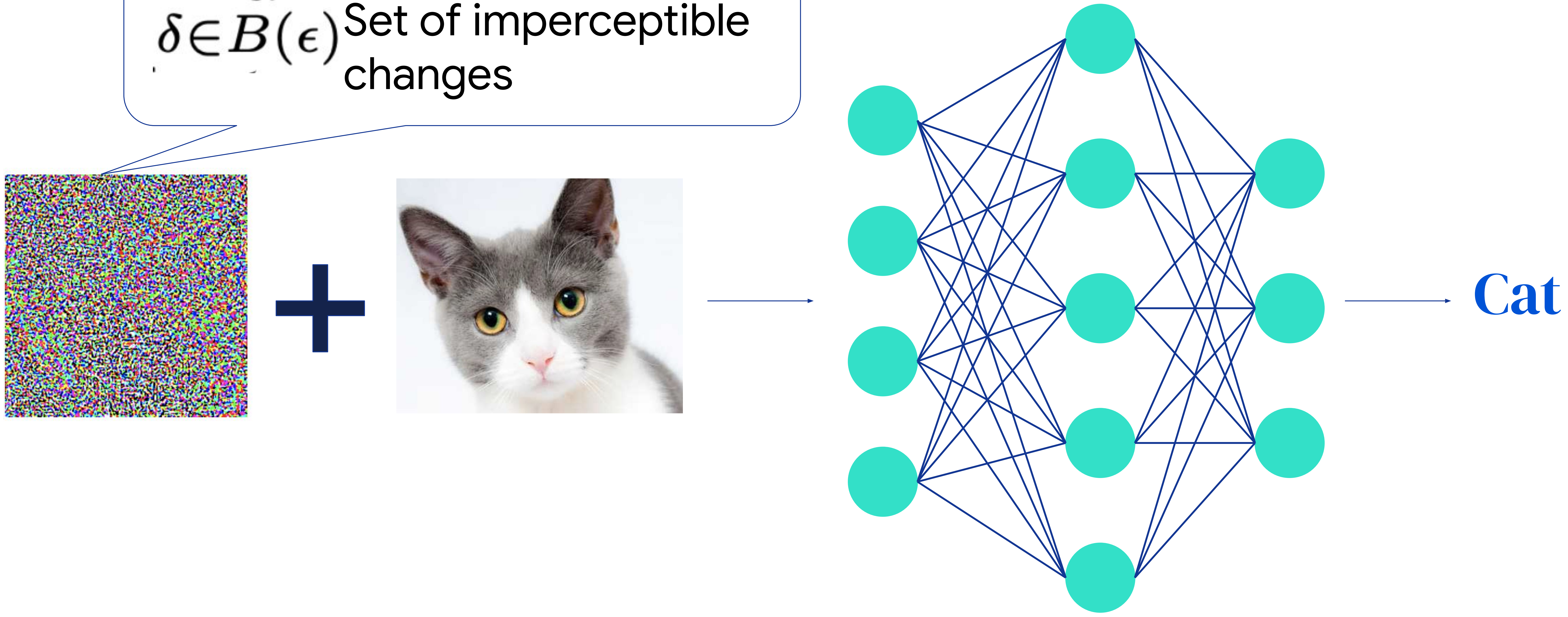
$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\underbrace{\ell(x; y, \theta)}_{\text{Cross entropy}} \right]$$

Adversarial Training

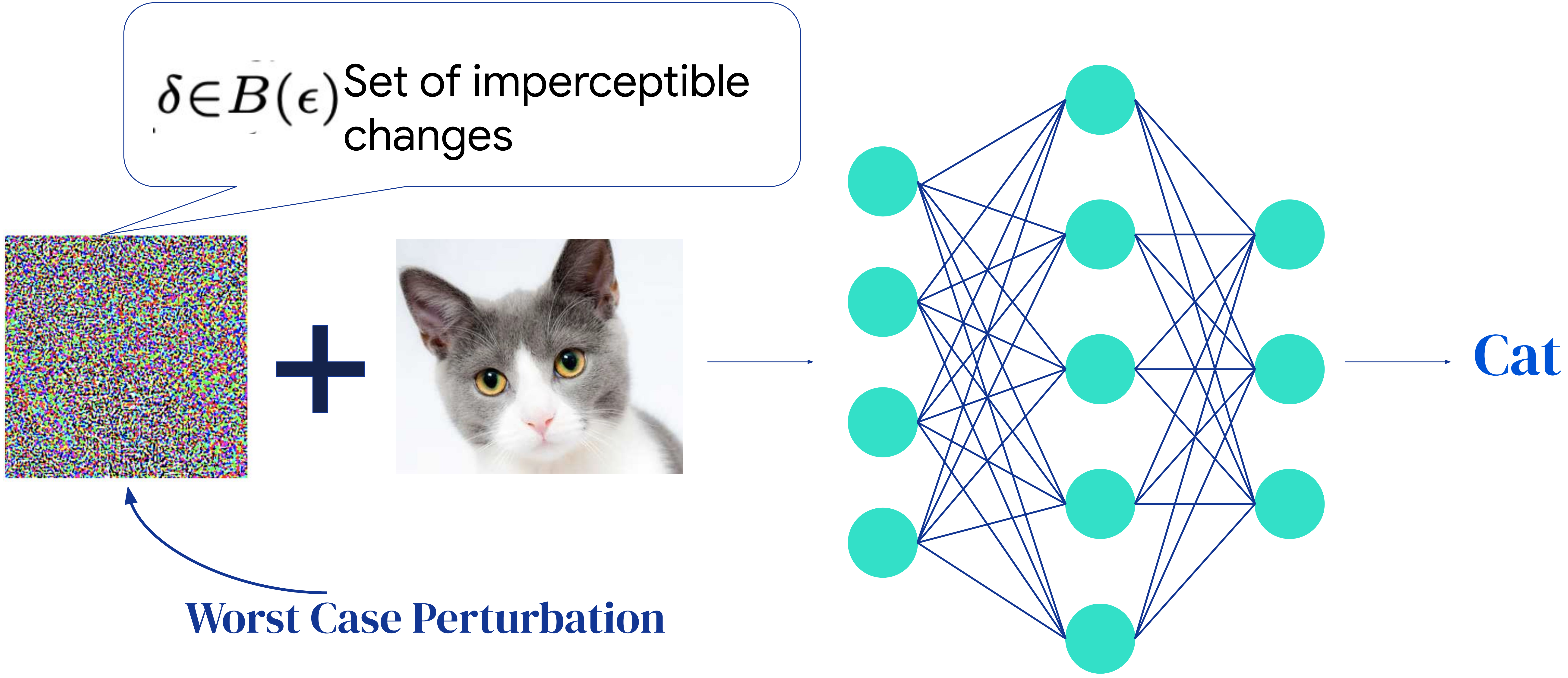


Adversarial Training

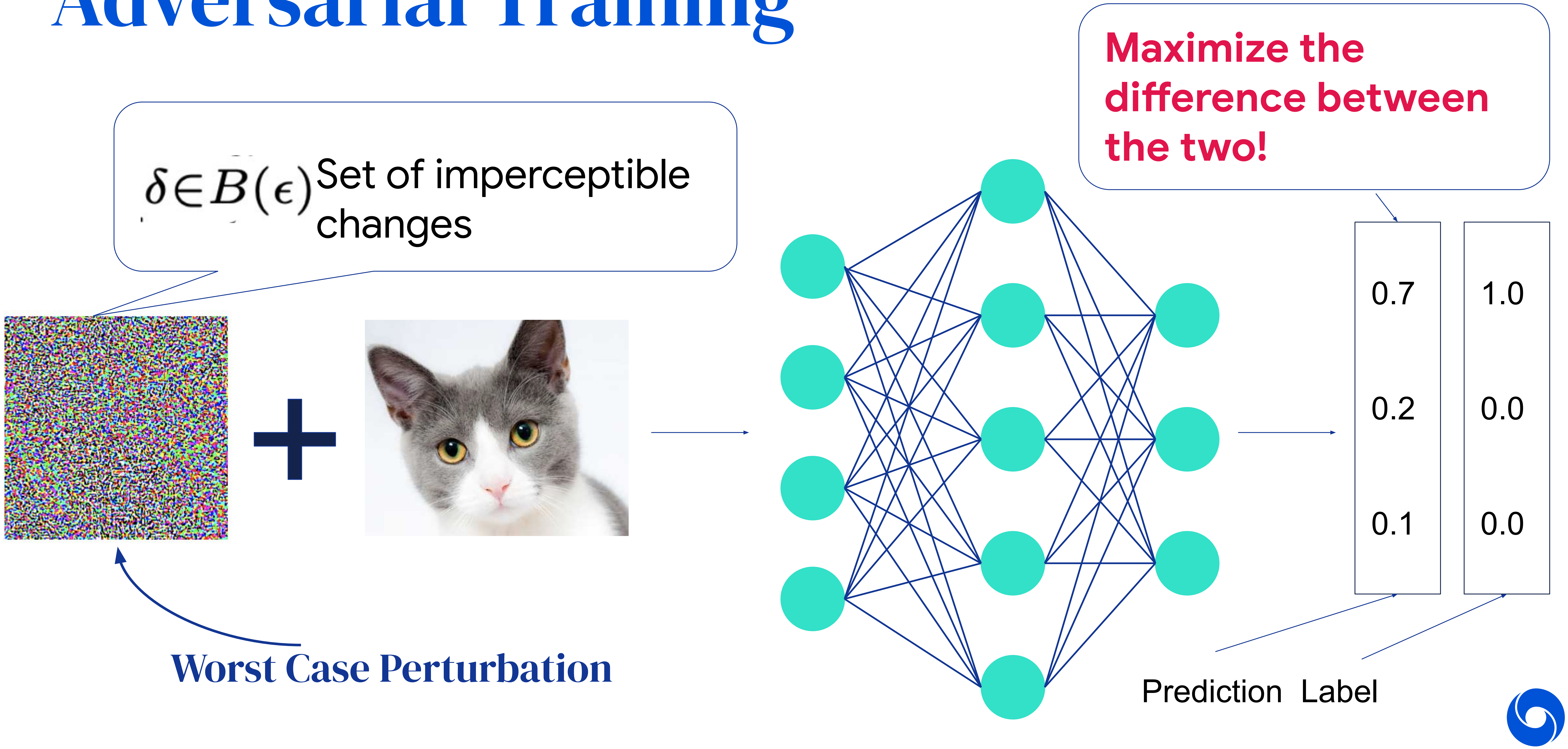
$\delta \in B(\epsilon)$ Set of imperceptible changes



Adversarial Training



Adversarial Training



Adversarial Training

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in B(\epsilon)} \ell(x+\delta; y, \theta) \right]$$

Set of Perturbations

Cross entropy

Perturbation

Adversarial Training

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in B(\epsilon)} \ell(x+\delta; y, \theta) \right]$$

Outer Minimization

Inner Maximization

DeepMind

3.1 Adversarial Evaluation: Finding the Worst Case

Adversarial Evaluation /Attacks

$$\max_{\delta \in B(\epsilon)} \ell(x + \delta; y, \theta)$$

- Find the worst case for each example in the test set.

- Then we evaluate the accuracy of this new test set, where each example is now replaced with the worst case adversarial image. This is known as *adversarial accuracy*.

Projected Gradient Ascent

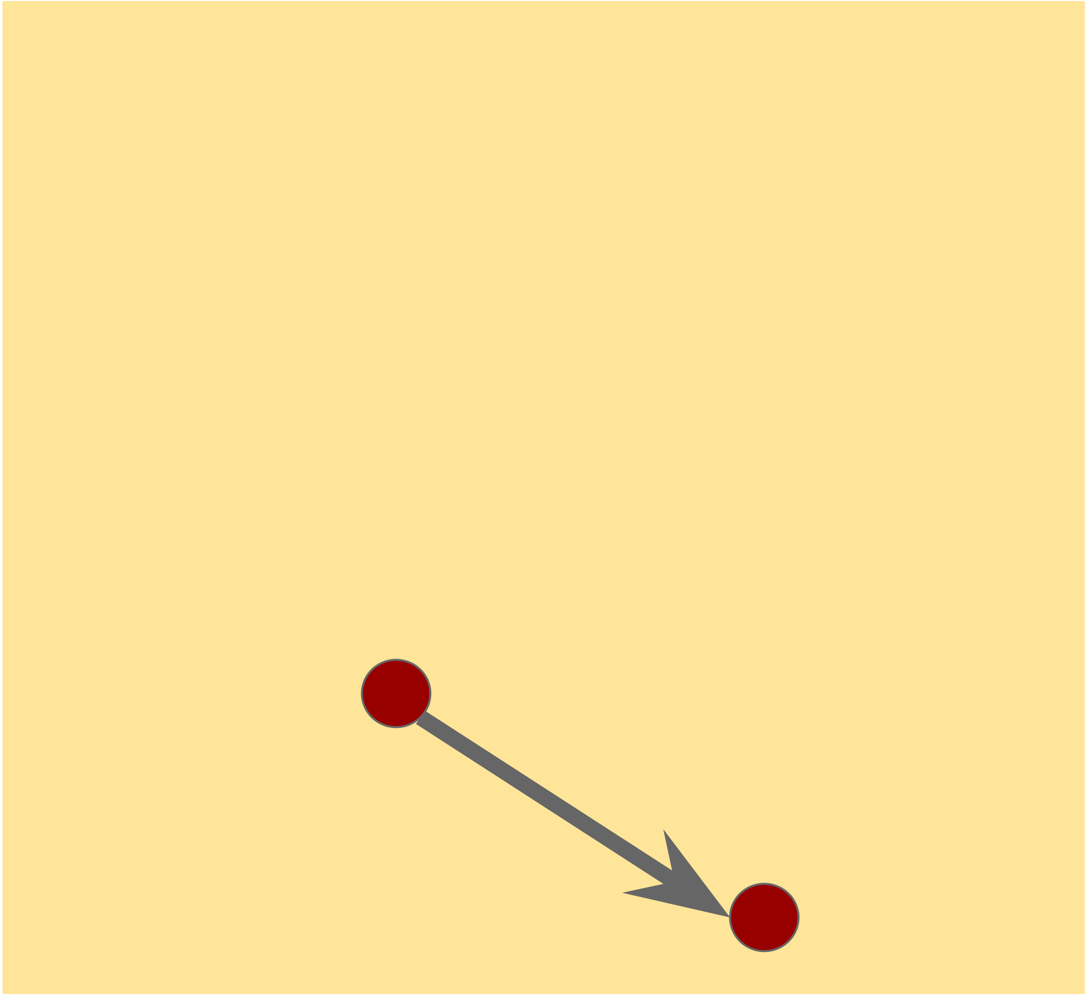
$$\max_{\delta \in B(\epsilon)} \ell(x + \delta; y, \theta)$$

Projected Gradient Ascent

$$\underbrace{\max_{\delta \in B(\epsilon)} \ell(x + \delta; y, \theta)}_{\text{Constrained Optimisation Problem}}$$

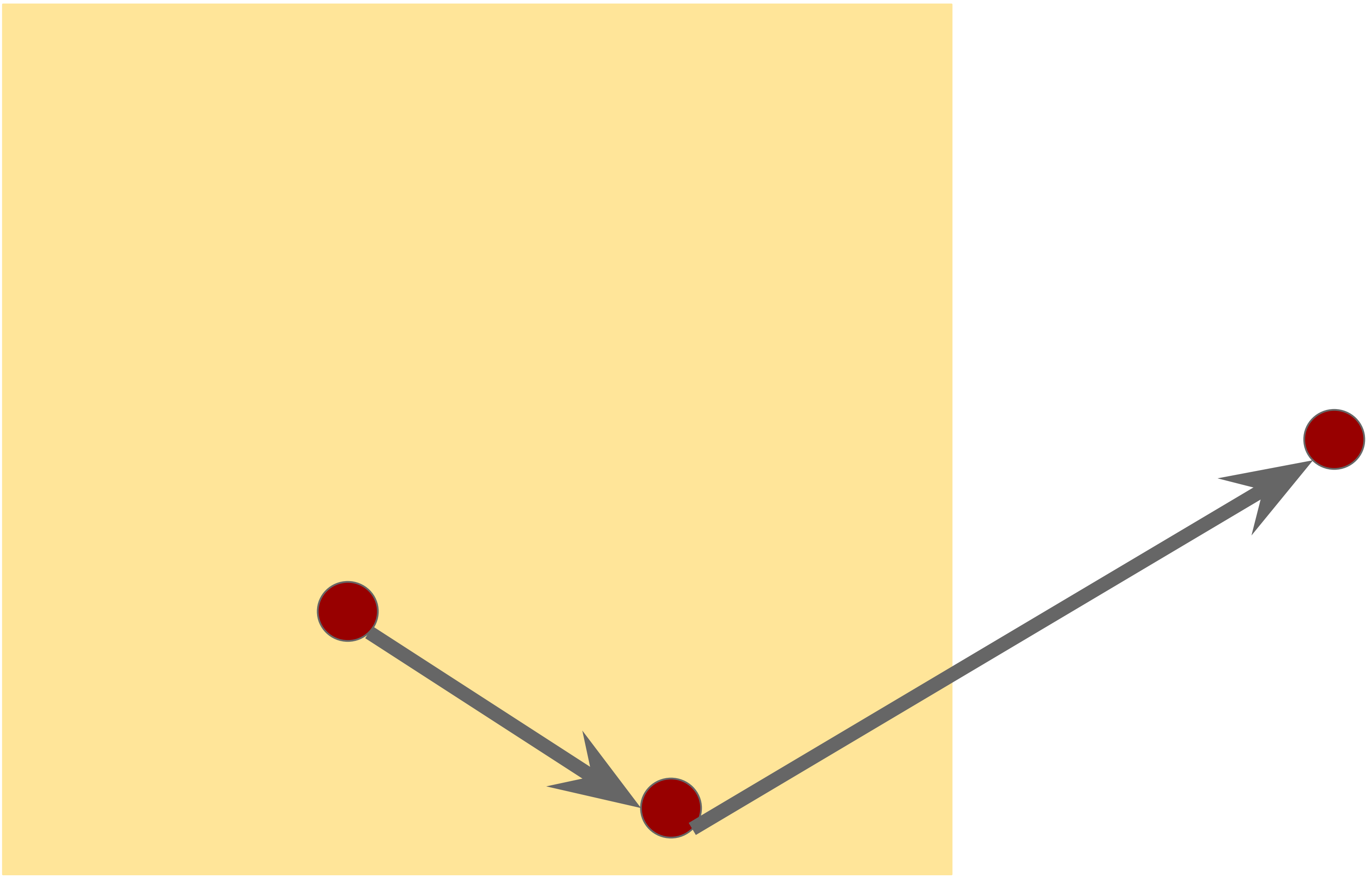
Projected Gradient Ascent

$$\max_{\delta \in B(\epsilon)} \ell(x + \delta; y, \theta)$$



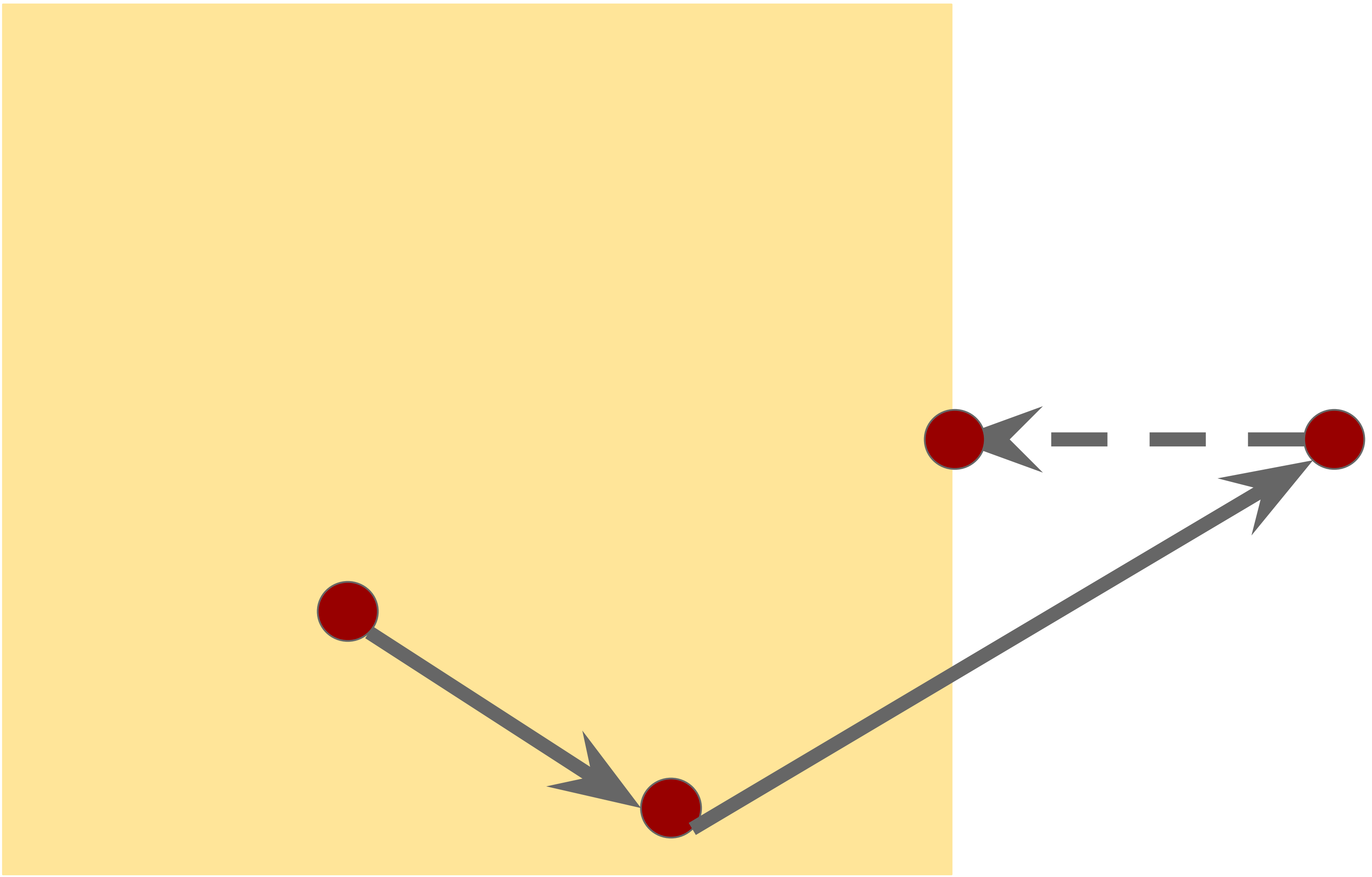
Projected Gradient Ascent

$$\max_{\delta \in B(\epsilon)} \ell(x + \delta; y, \theta)$$



Projected Gradient Ascent

$$\max_{\delta \in B(\epsilon)} \ell(x + \delta; y, \theta)$$



Projected Gradient Ascent

$$\max_{\delta \in B(\epsilon)} \ell(x + \delta; y, \theta)$$

Update Step: $\delta \leftarrow \text{Proj}\left(\delta + \eta \nabla_\delta \ell(x + \delta, y; \theta)\right)$

$$\text{Proj}(\delta) = \text{argmin}_{\delta' \in B(\epsilon)} \parallel \delta - \delta' \parallel$$

Fast Gradient Sign Method (Iterated)

$$\delta \leftarrow \mathrm{Proj}\left(\delta + \eta \nabla_{\delta} \ell(x + \delta, y; \theta)\right)$$

$$\downarrow$$

$$\delta \leftarrow \mathrm{Proj}\left(\delta + \eta \,\mathrm{sgn}(\nabla_{\delta} \ell(x + \delta, y; \theta))\right)$$

Any Optimiser can be used

$$\delta \leftarrow \mathrm{Proj}\left(\delta + \eta \nabla_\delta \ell(x + \delta, y; \theta)\right)$$

$$\downarrow$$

$$\delta \leftarrow \mathrm{Proj}\left(\delta + \eta \mathrm{Opt}(\nabla_\delta \ell(x + \delta, y; \theta))\right)$$

Note that we can replace the gradient by any alterations made by an optimiser, such as momentum optimisation or Adam optimisation.

“Strengths” of Adversarial Evaluation

- Adversarial accuracy is *dependent* on your choice of evaluation.
- Stronger adversarial evaluation should give the lower adversarial accuracy.
- We should *always* try to evaluate our network such that we can obtain the *lowest* adversarial accuracy.
- Strength of your evaluation depends on a few heuristics:
  - Number of steps of projected gradient ascent (PGA).
  - Number of random initialisations of perturbations.
  - The optimiser used.
  - Black box adversarial evaluation e.g. Square Attack

Dangers of Weak Adversarial Evaluation

Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

Anish Athalye [* 1] Nicholas Carlini [* 2] David Wagner [2]

Adversarial Risk and the Dangers of Evaluating Against Weak Attacks

Jonathan Uesato [1] Brendan O'Donoghue [1] Aaron van den Oord [1] Pushmeet Kohli [1]

Dangers of Weak Adversarial Evaluation

Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

Advers **inst Weak Attacks**

J **Pushmeet Kohli** [1]

Defense	Dataset	Distance	Accuracy
Buckman et al. (2018)	CIFAR	0.031 (ℓ_∞)	0%*
Ma et al. (2018)	CIFAR	0.031 (ℓ_∞)	5%
Guo et al. (2018)	ImageNet	0.005 (ℓ_2)	0%*
Dhillon et al. (2018)	CIFAR	0.031 (ℓ_∞)	0%
Xie et al. (2018)	ImageNet	0.031 (ℓ_∞)	0%*
Song et al. (2018)	CIFAR	0.031 (ℓ_∞)	9%*
Samangouei et al. (2018)	MNIST	0.005 (ℓ_2)	55%**
Madry et al. (2018)	CIFAR	0.031 (ℓ_∞)	47%
Na et al. (2018)	CIFAR	0.015 (ℓ_∞)	15%

Stronger Adversarial Evaluation Gives Better Evaluation of Progress

Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks

Francesco Croce [1] Matthias Hein [1]

Stronger Adversarial Evaluation Gives Better Evaluation of Progress

#	paper	model	clean	report.	AA	AA+
1	(Carmon et al., 2019)‡	*available*	89.69	62.5	59.65	59.50
2	(Sehwag et al., 2020)‡	*available*	88.98	-	57.24	57.11
3	(Wang et al., 2020)‡	*available*	87.50	65.04	56.69	56.26
4	(Alayrac et al., 2019)‡	*available*	86.46	56.30	56.92	56.01
⋮						
21	(Zhang & Xu, 2020)	*available*	90.25	68.7	38.57	
22	(Kim & Wang, 2020)	*available*	91.51	57.23	36.10	
23	(Jang et al., 2019)	*available*	78.91	37.40	35.09	
24	(Wang & Zhang, 2019)	*available*	92.80	58.6	30.96	
25	(Xiao et al., 2020)*	*available*	79.28	52.4	17.99	
26	(Jin & Rinard, 2020)	*available*	90.84	71.22	4.61	
27	(Mustafa et al., 2019)	*available*	89.16	32.32	0.55	
28	(Chan et al., 2020)	*retrained*	93.79	15.5	0.18	

Stronger Adversarial Evaluation Gives Better Evaluation of Progress

#	paper	model	clean	report.	AA	AA+
1	(Carmon et al., 2019)‡	*available*	89.69	62.5	59.65	59.50
2	(Sehwag et al., 2020)‡	*available*	88.98	-	57.24	57.11
3	(Wang et al., 2020)‡	*available*	87.50	65.04	56.69	56.26
4	(Alayrac et al., 2019)‡	*available*	86.46	56.30	56.92	56.01
⋮						
21	(Zhang & Xu, 2020)	*available*	90.25	68.7	38.57	
22	(Kim & Wang, 2020)	*available*	91.51	57.23	36.10	
23	(Jang et al., 2019)	*available*	78.91	37.40	35.09	
24	(Wang & Zhang, 2019)	*available*	92.80	58.6	30.96	
25	(Xiao et al., 2020)*	*available*	79.28	52.4	17.99	
26	(Jin & Rinard, 2020)	*available*	90.84	71.22	4.61	
27	(Mustafa et al., 2019)	*available*	89.16	32.32	0.55	
28	(Chan et al., 2020)	*retrained*	93.79	15.5	0.18	

Stronger Adversarial Evaluation Gives Better Evaluation of Progress

#	paper	model	clean	report.	AA	AA+
1	(Carmon et al., 2019)‡	*available*	89.69	62.5	59.65	59.50
2	(Sehwag et al., 2020)‡	*available*	88.98	-	57.24	57.11
3	(Wang et al., 2020)‡	*available*	87.50	65.04	56.69	56.26
4	(Alayrac et al., 2019)‡	*available*	86.46	56.30	56.92	56.01
⋮						
21	(Zhang & Xu, 2020)	*available*	90.25	68.7	38.57	
22	(Kim & Wang, 2020)	*available*	91.51	57.23	36.10	
23	(Jang et al., 2019)	*available*	78.91	37.40	35.09	
24	(Wang & Zhang, 2019)	*available*	92.80	58.6	30.96	
25	(Xiao et al., 2020)*	*available*	79.28	52.4	17.99	
26	(Jin & Rinard, 2020)	*available*	90.84	71.22	4.61	
27	(Mustafa et al., 2019)	*available*	89.16	32.32	0.55	
28	(Chan et al., 2020)	*retrained*	93.79	15.5	0.18	

DeepMind

3.2 Gradient Obfuscation

Adversarial Training

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in B(\epsilon)} \ell(x+\delta; y, \theta) \right]$$

Outer Minimization

Inner Maximization

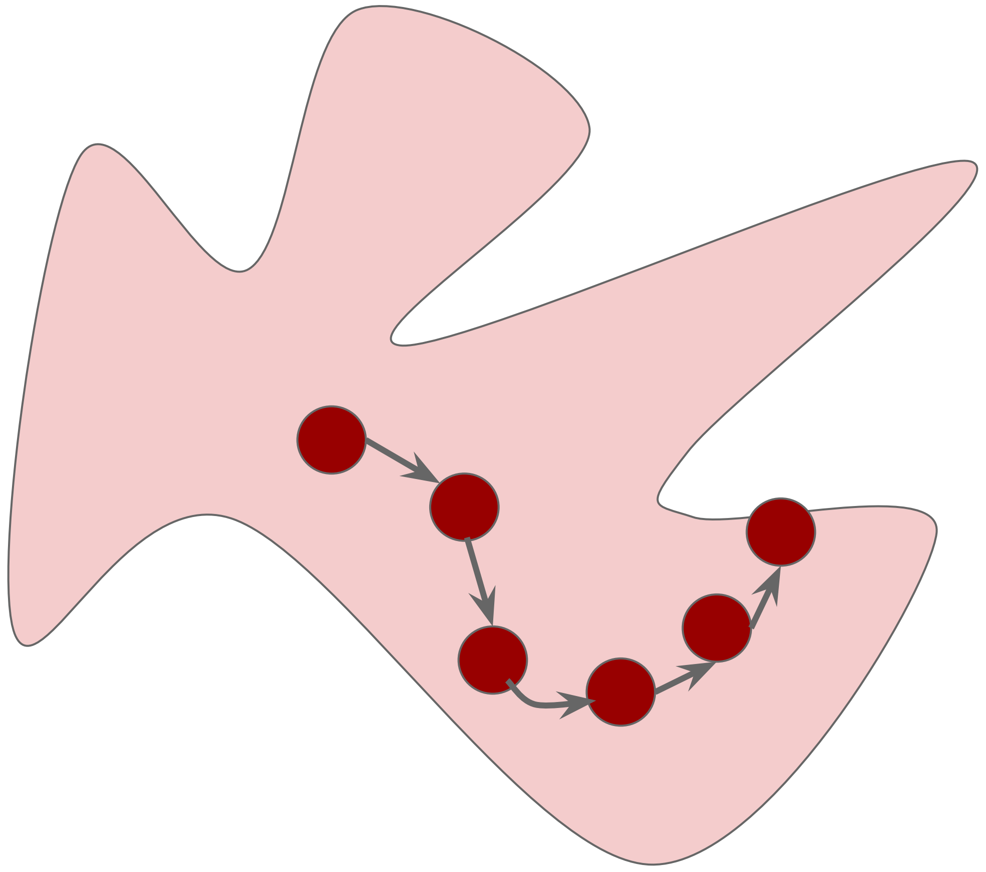
Adversarial Training

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in B(\epsilon)} \ell(x+\delta; y, \theta) \right]$$

Inner Maximization

Adversarial Training

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in B(\epsilon)} \ell(x+\delta; y, \theta) \right]$$



Adversarial Training made Cheaper

$$\max_{\delta \in B(\epsilon)} \ell(x+\delta; y, \theta)$$

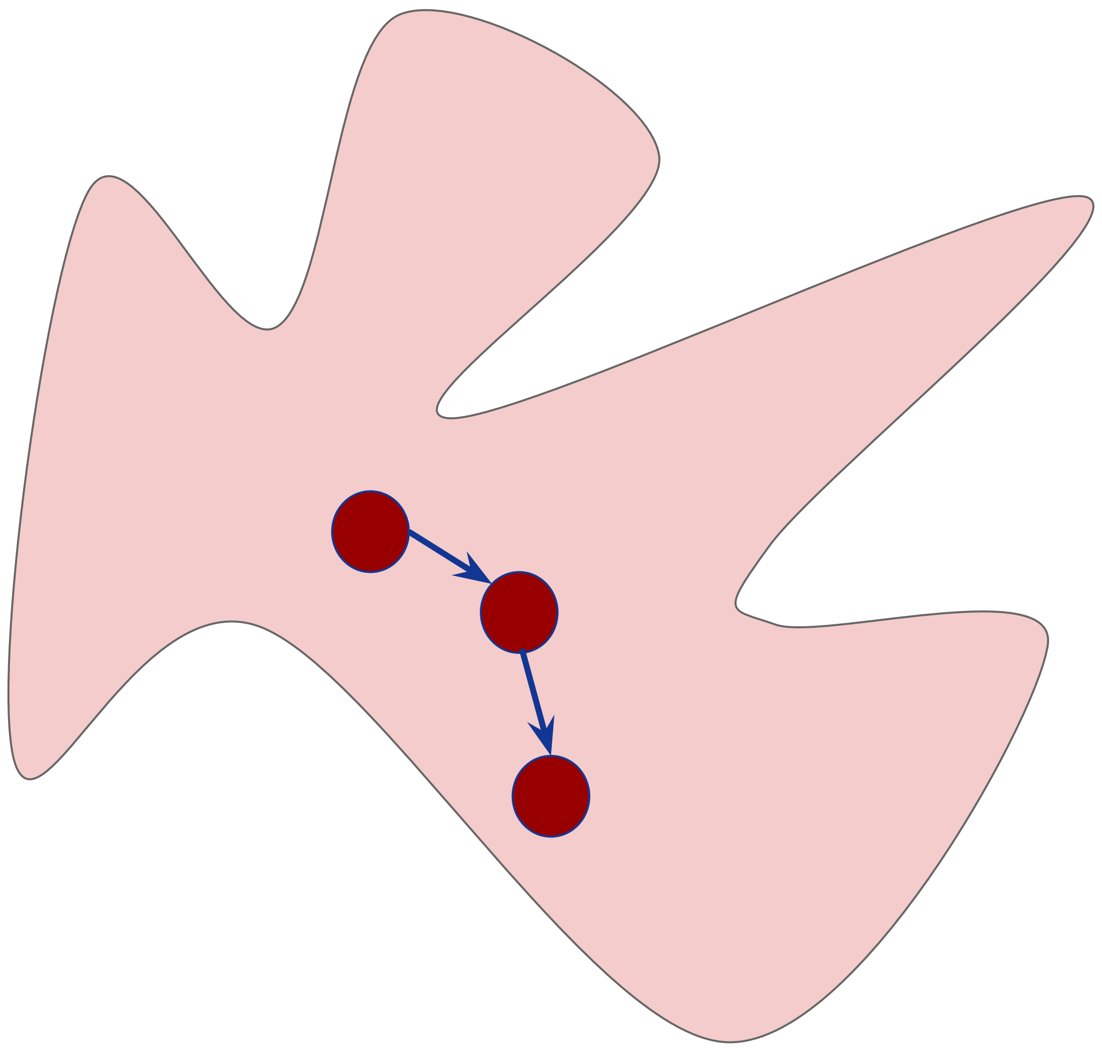
Adversarial Training made Cheaper

$$\max_{\delta \in B(\epsilon)} \ell(x + \delta; y, \theta)$$

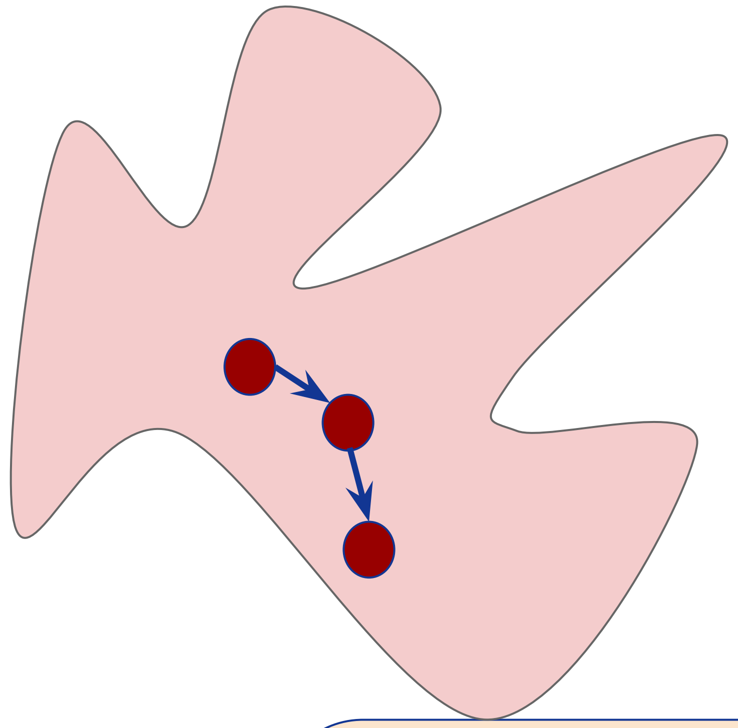
Cheaper Adversarial Training :
Few steps of Gradient Ascent for
Inner Maximization.

Adversarial Training made Cheaper

$$\max_{\delta \in B(\epsilon)} \ell(x+\delta; y, \theta)$$



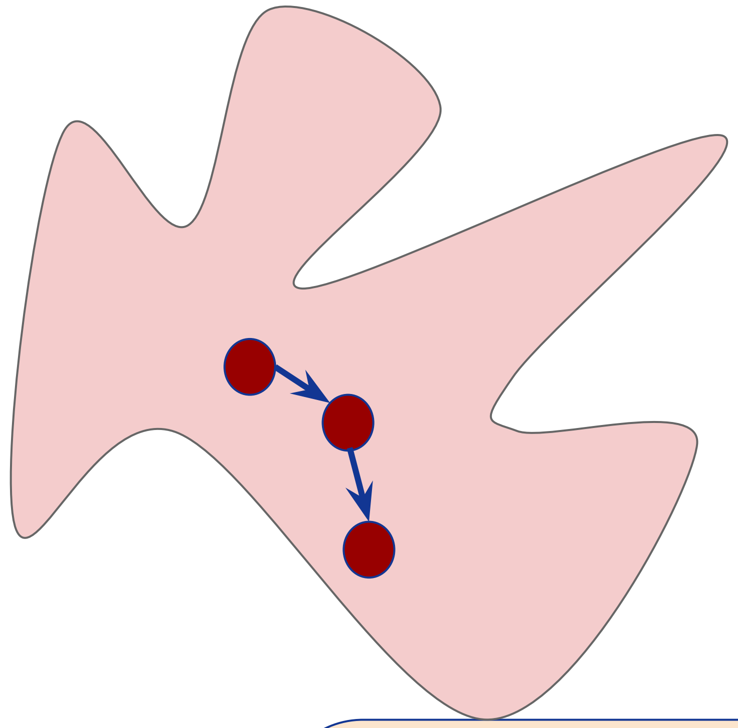
Cheaper Adversarial Training :
Few steps of Gradient Ascent for Inner Maximization.



Gradient Obfuscation

$$\max_{\delta \in B(\epsilon)} \ell(x + \delta; y, \theta)$$

RESULT:
The **network learns to cheat** by making a **highly non-linear surface** such that the **optimum is hard to find** using a suboptimal optimization procedure.

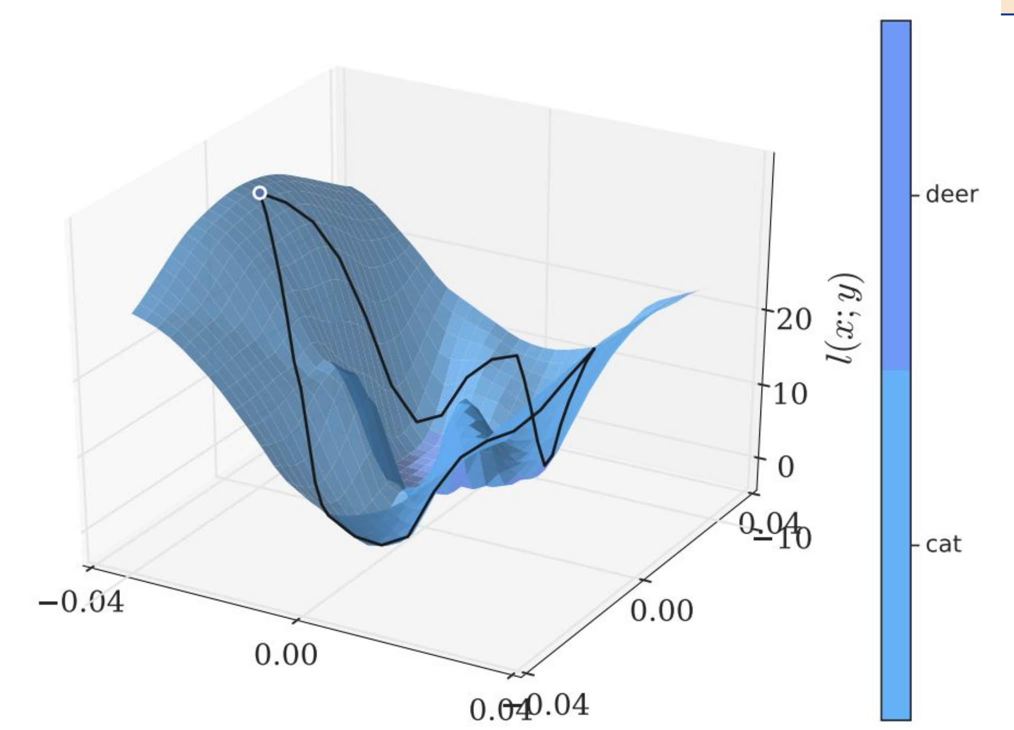


Gradient Obfuscation

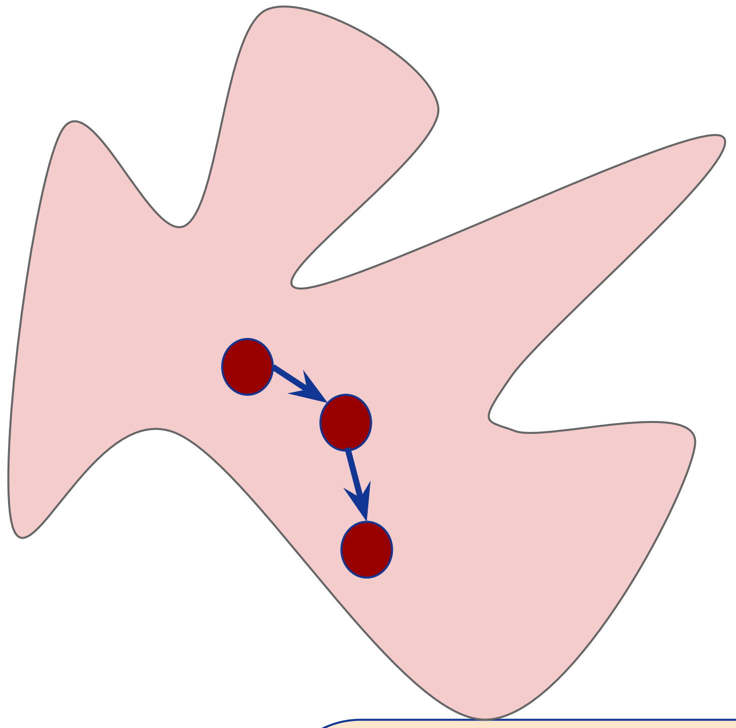
$$\max_{\delta \in B(\epsilon)} \ell(x + \delta; y, \theta)$$

RESULT:
The **network learns to cheat** by making a **highly non-linear surface** such that the **optimum is hard to find** using a suboptimal optimization procedure.

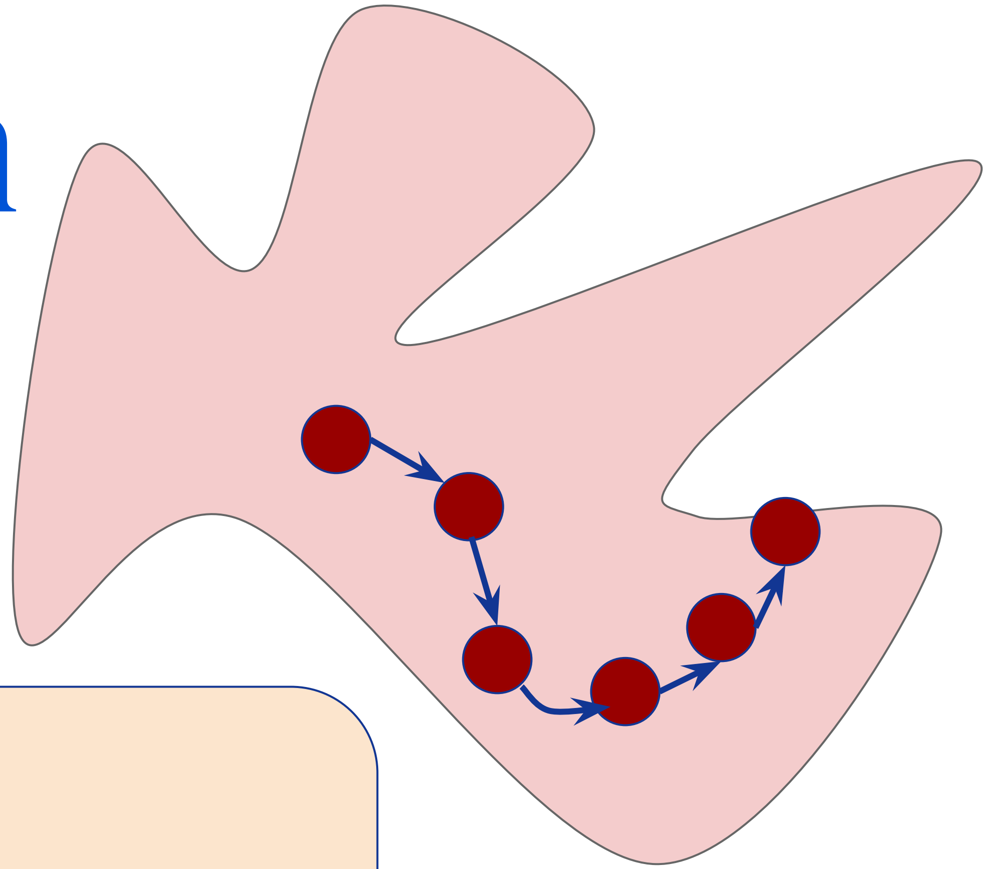
Loss plotted for two input dimensions



Gradient Obfuscation



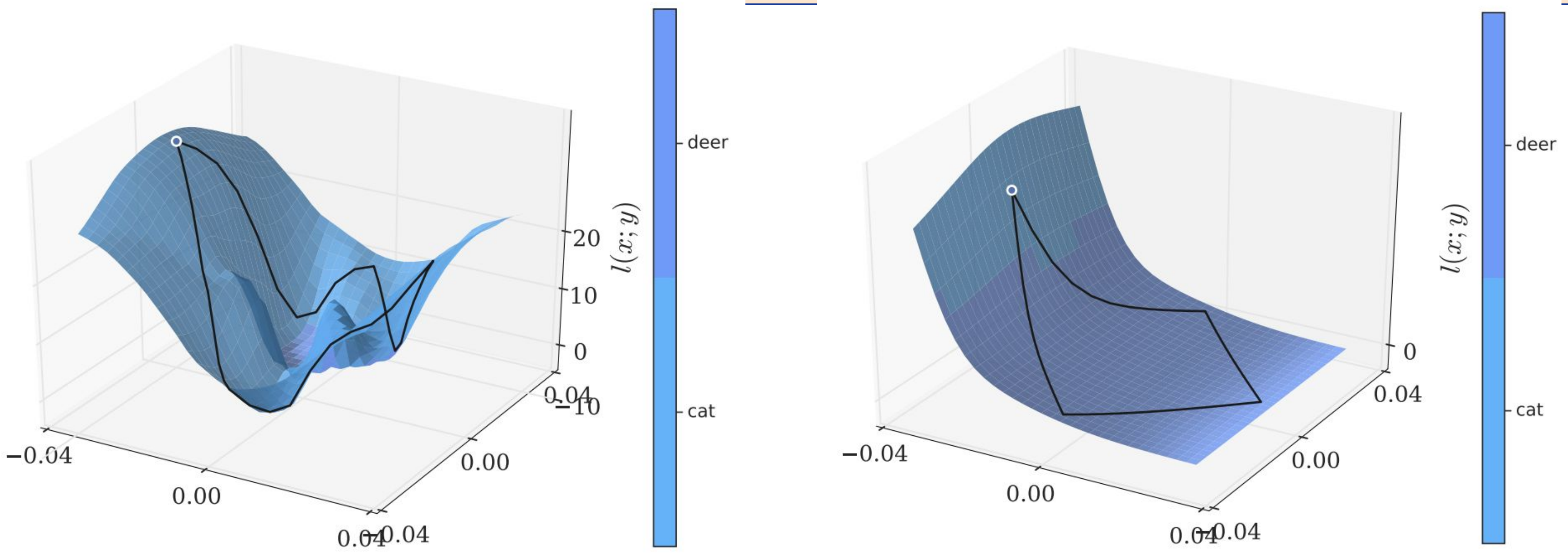
$$\max_{\delta \in B(\epsilon)} \ell(x + \delta; y, \theta)$$



RESULT:

The **network learns to cheat** by making a **highly non-linear surface** such that the **optimum is hard to find** using a suboptimal optimization procedure.

Loss plotted for two input dimensions



DeepMind

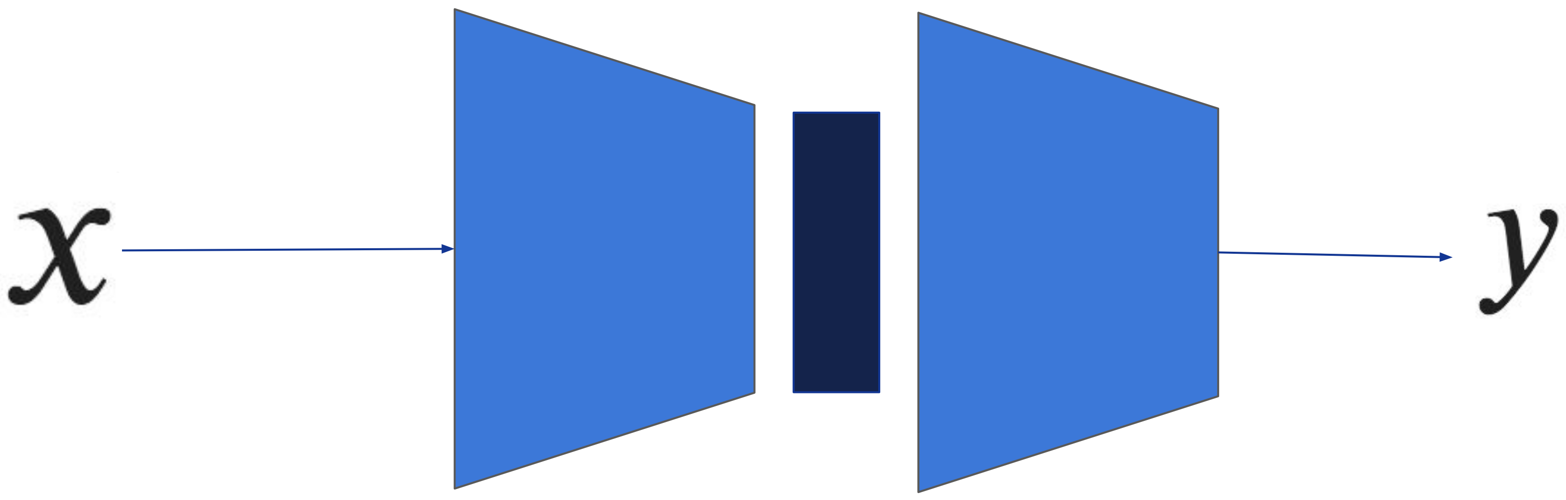
3.3 Verification Algorithms

Verification Algorithms

- **Complete**
  - Exhaustive Proof
- **Incomplete**
  - If proof can be found then it is a guarantee but proof cannot always be found.

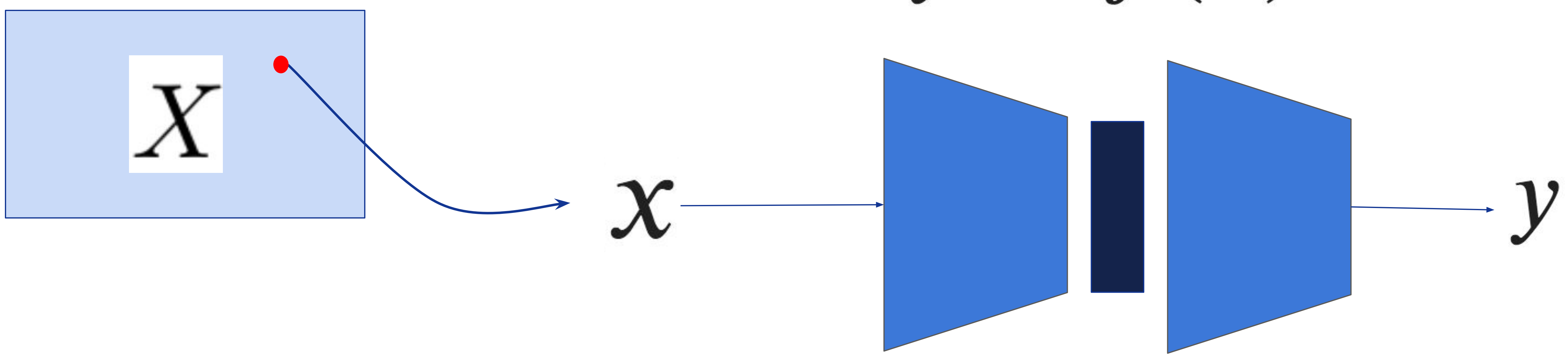
Verification of Specifications

$$y = f(x)$$

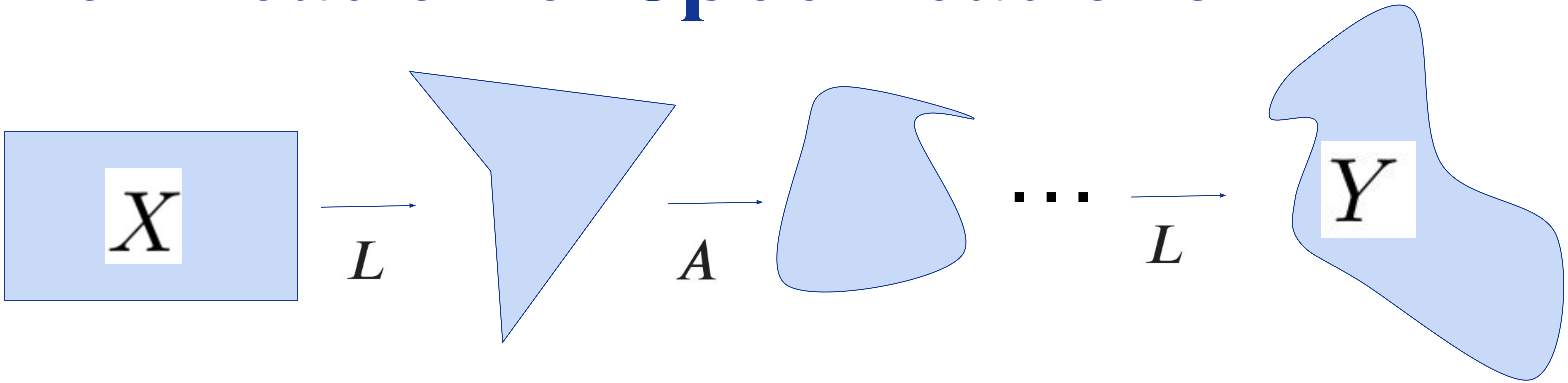


Verification of Specifications

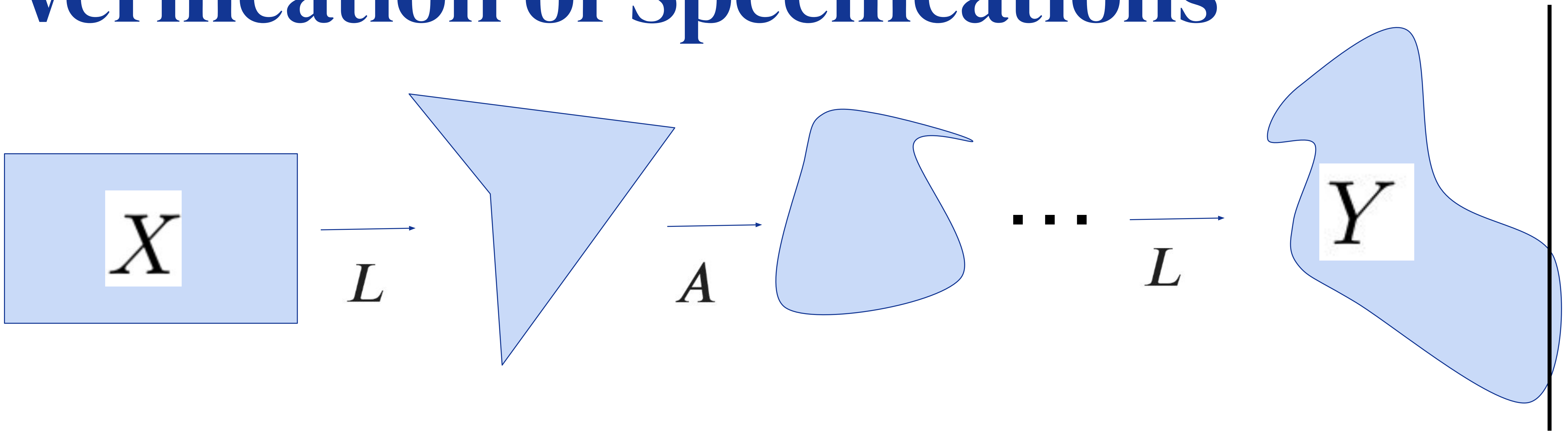
$$y = f(x)$$



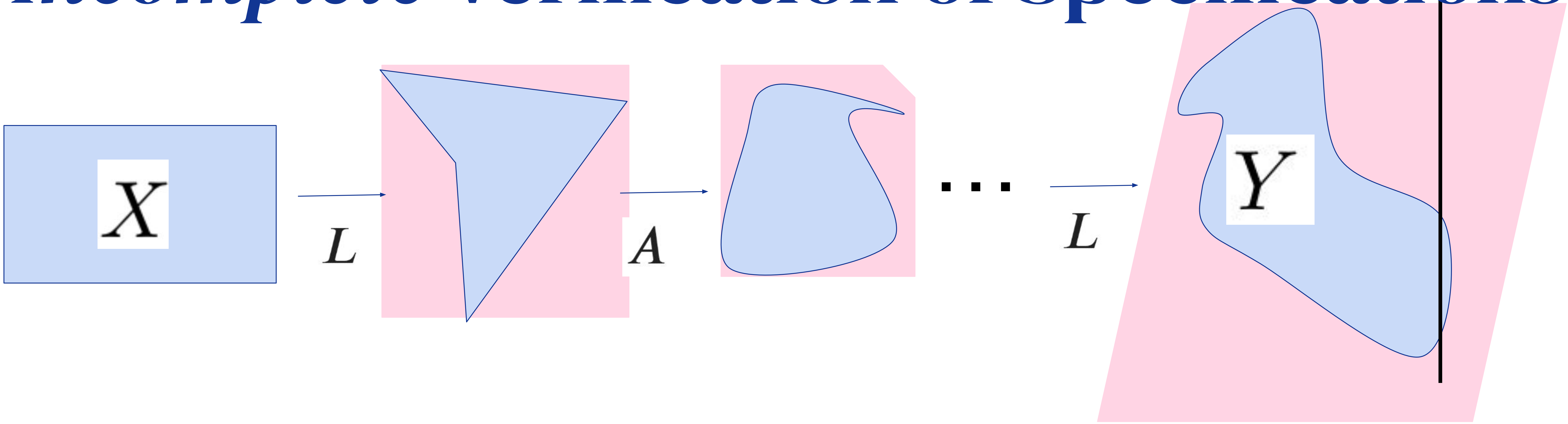
Verification of Specifications



Verification of Specifications

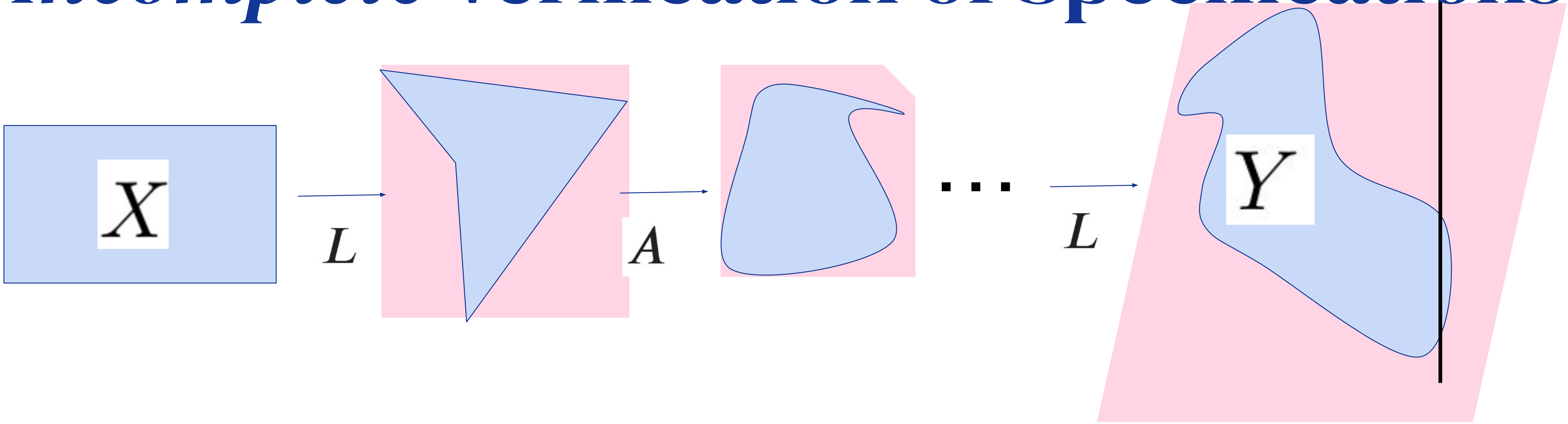


Incomplete Verification of Specifications



$$\underline{x}^0 \leq x \leq \overline{x}^0$$

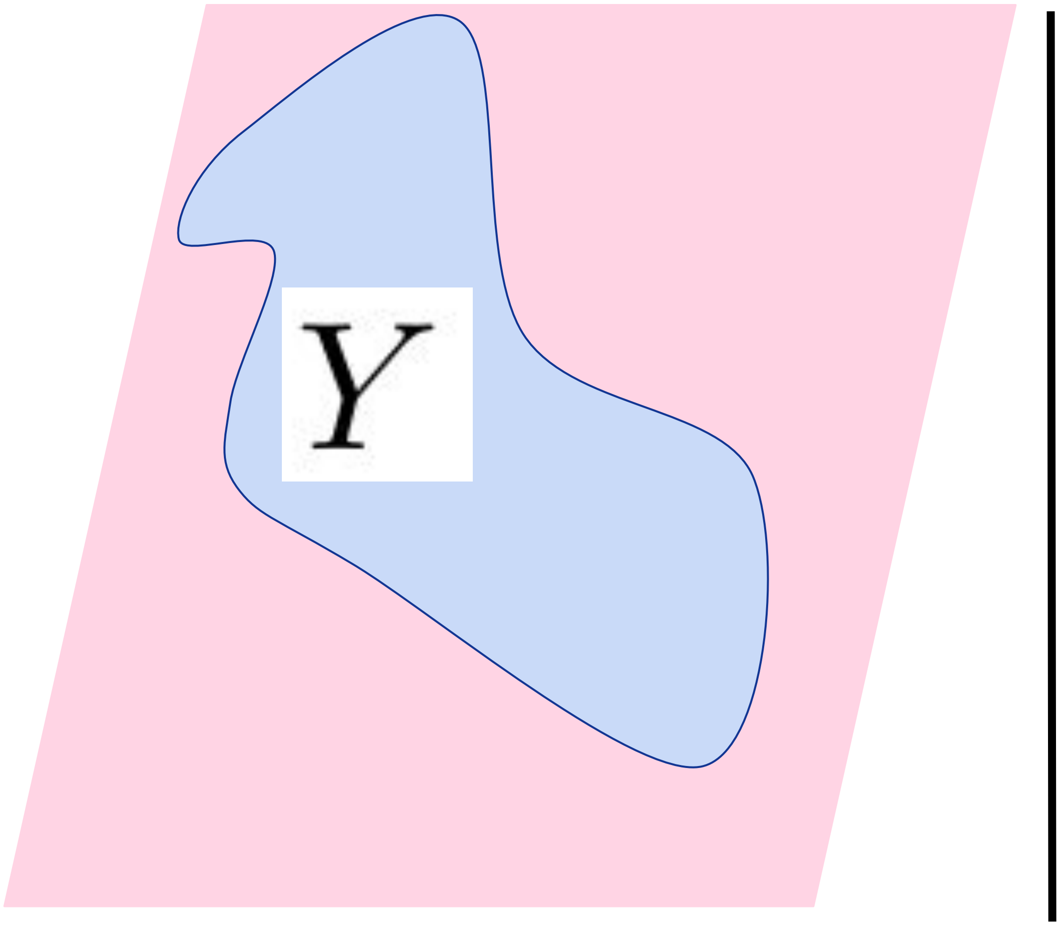
Incomplete Verification of Specifications



$$\underline{x}^0 \leq x \leq \overline{x}^0$$

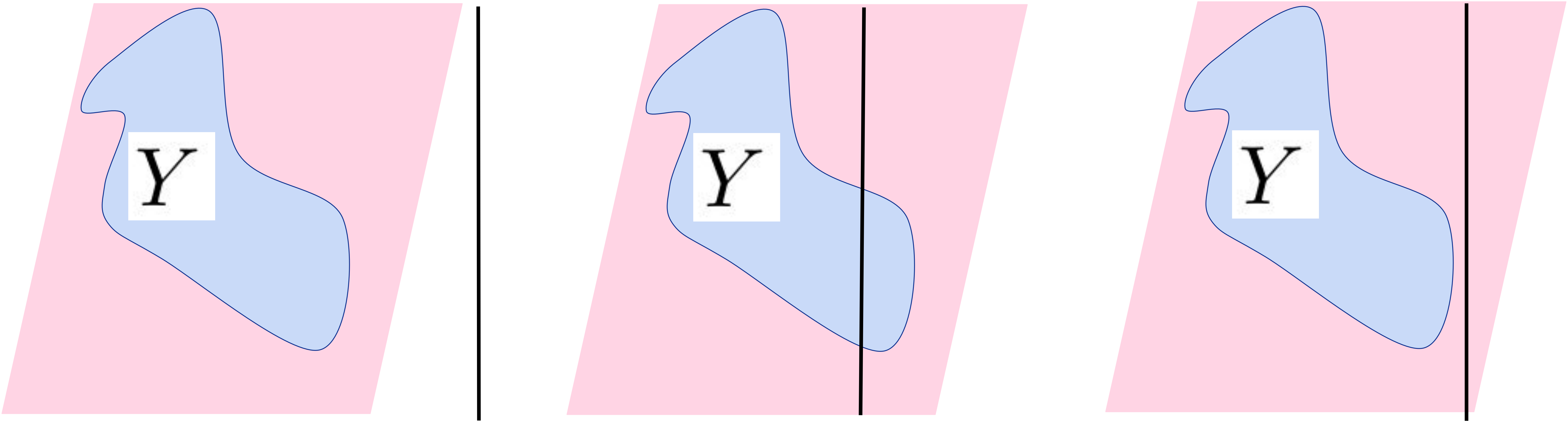
$$\underline{z}^l = [W^l]_+ \underline{x}^l + [W^l]_- \overline{x}^l + b^l$$
$$\overline{z}^l = [W^l]_+ \overline{x}^l + [W^l]_- \underline{z}^l + b^l$$
$$\underline{x}^{l+1} = h^l\left(\underline{z}^l\right)$$
$$\overline{x}^{l+1} = h^l\left(\overline{z}^l\right)$$

Incomplete Verification of Specifications



Y satisfies the specification

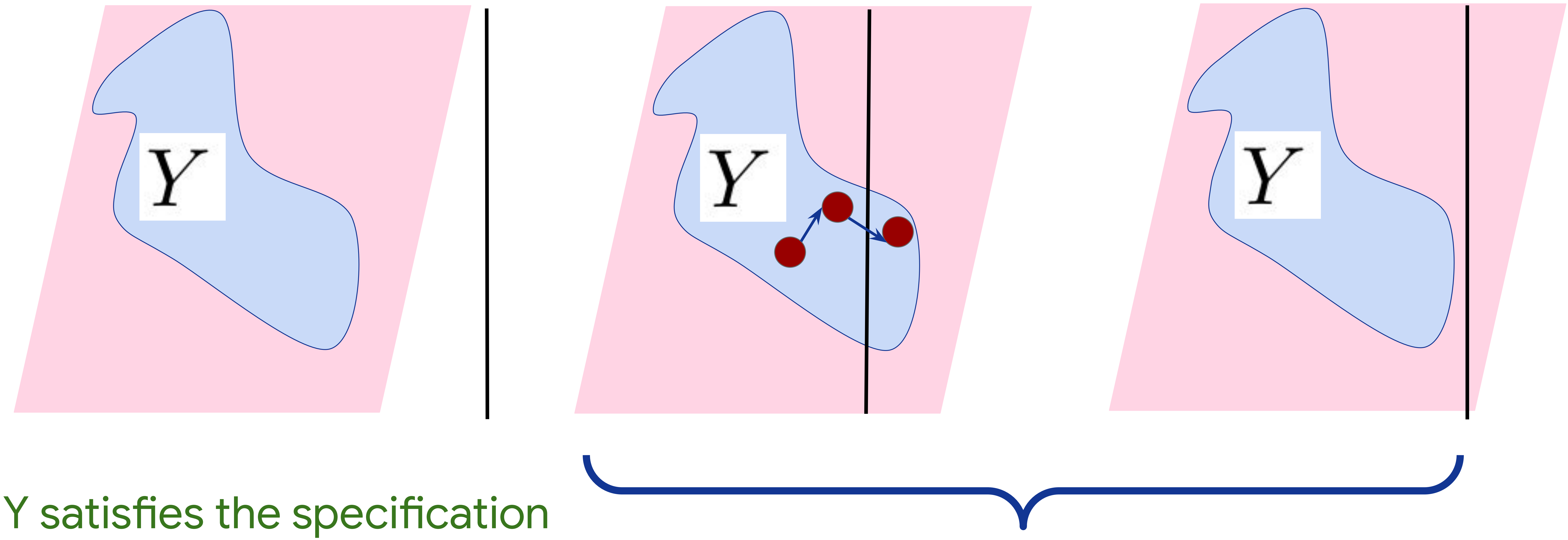
Incomplete Verification of Specifications



Y satisfies the specification

Nothing can said about Y

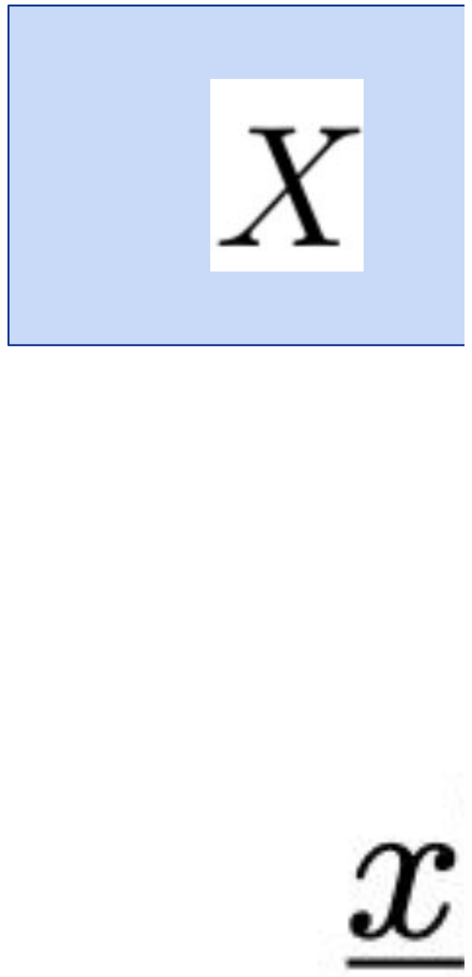
Incomplete Verification of Specifications



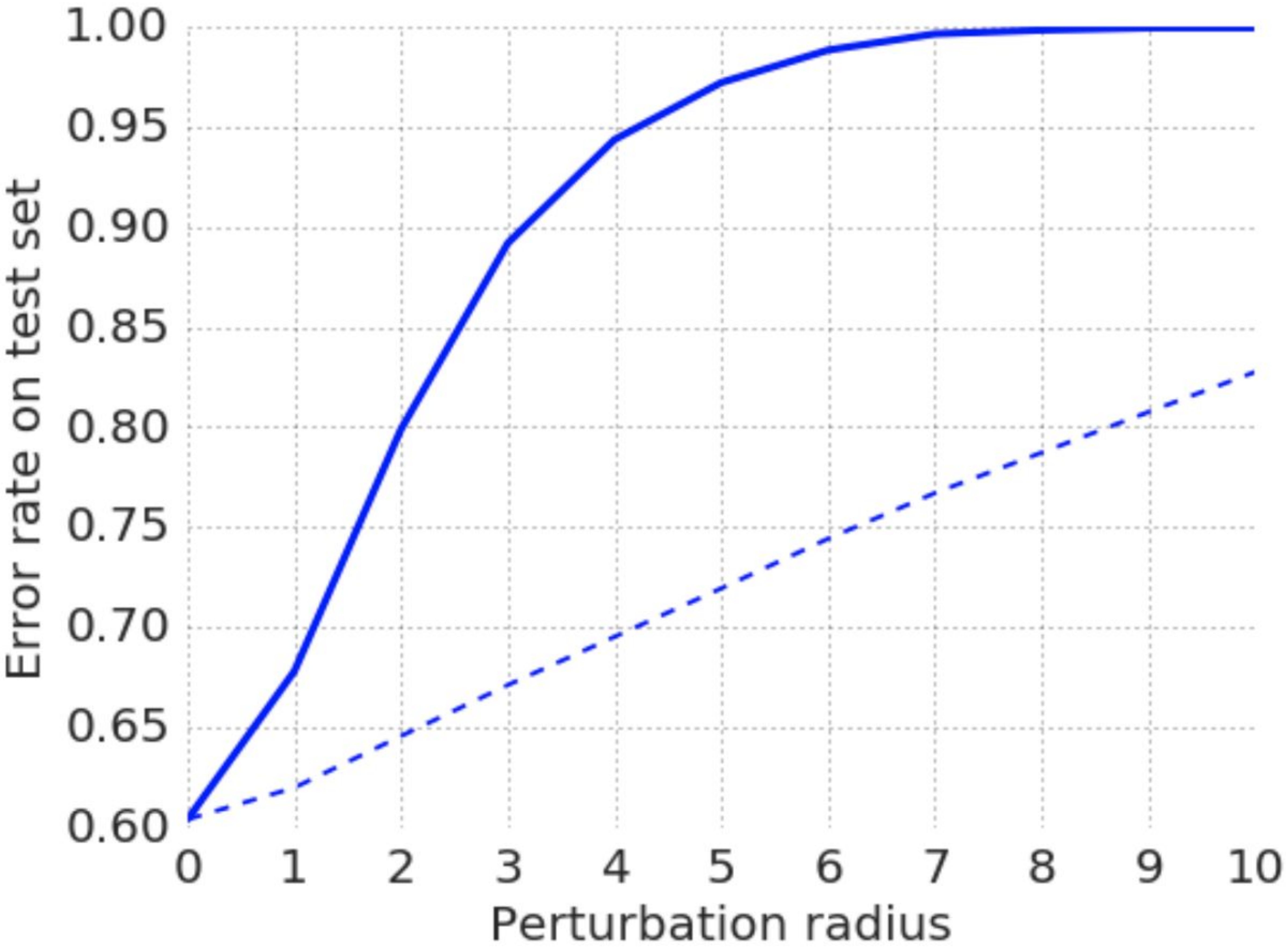
Y satisfies the specification

Nothing can said about Y

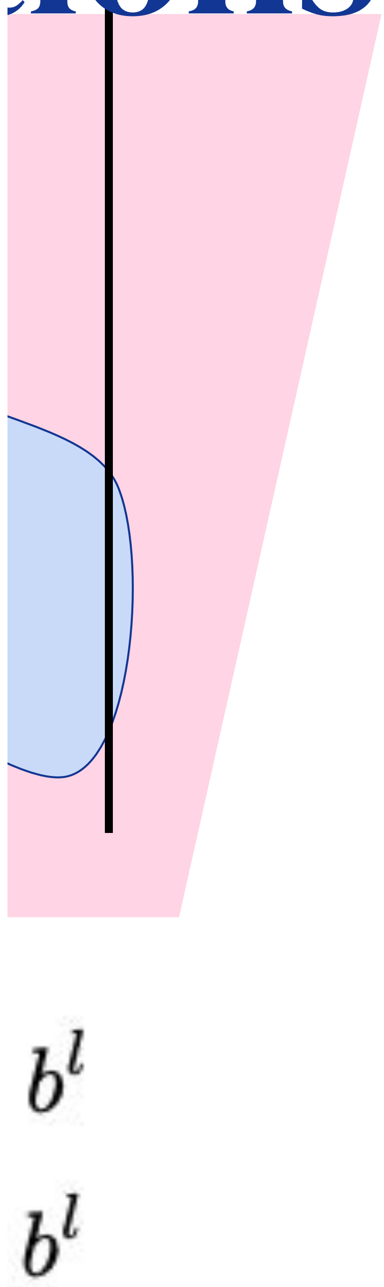
Incon … *ions*



X
x



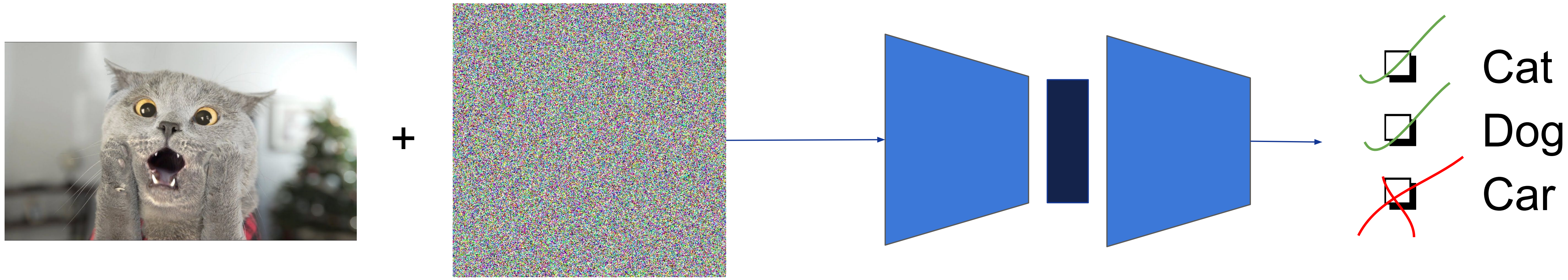
Error rate on test set
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0 1 2 3 4 5 6 7 8 9 10
Perturbation radius



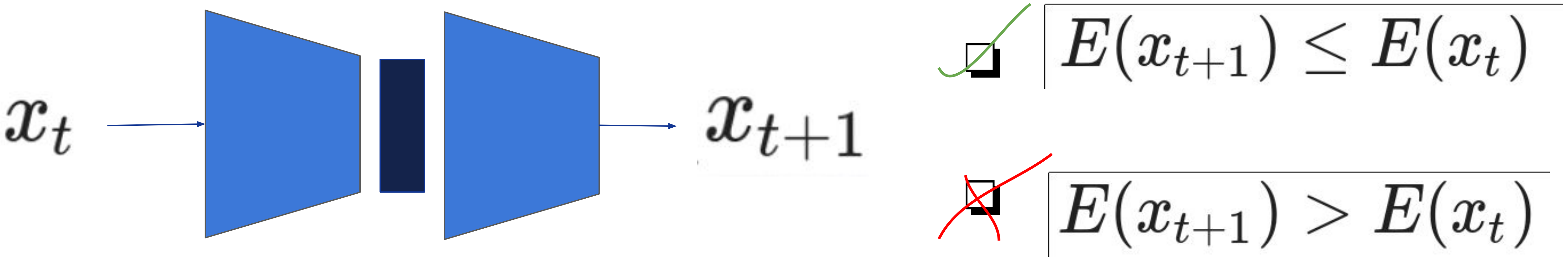
b^l
b^l

Other Specifications

Semantic Consistency:



Physics Consistency :



4 Ethics and Technology

Ethics & Machine Learning

Credit: Michal Lozma

Training Data

Training Data

- Consent

Training Data

- Consent
- Representation

Training Data

- Consent
- Representation
- Prejudicial Associations

Algorithmic Bias

BuzzFeed News
REPORTING TO YOU

TECH

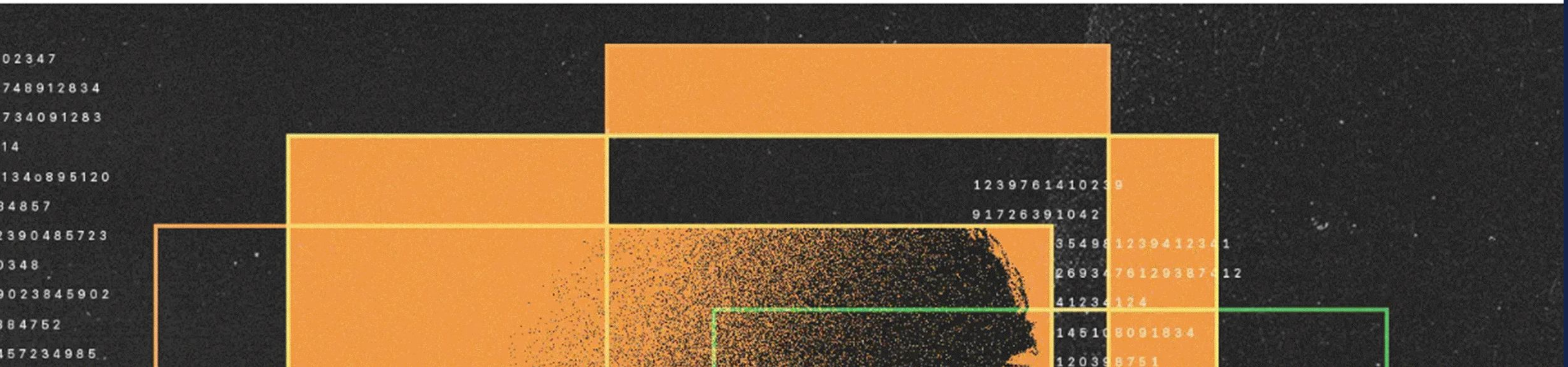
India Is Creating A National Facial Recognition System, And Critics Are Afraid Of What Will Happen Next

“Unless we all get plastic surgery at the same time, there’s nothing we can do about it.”

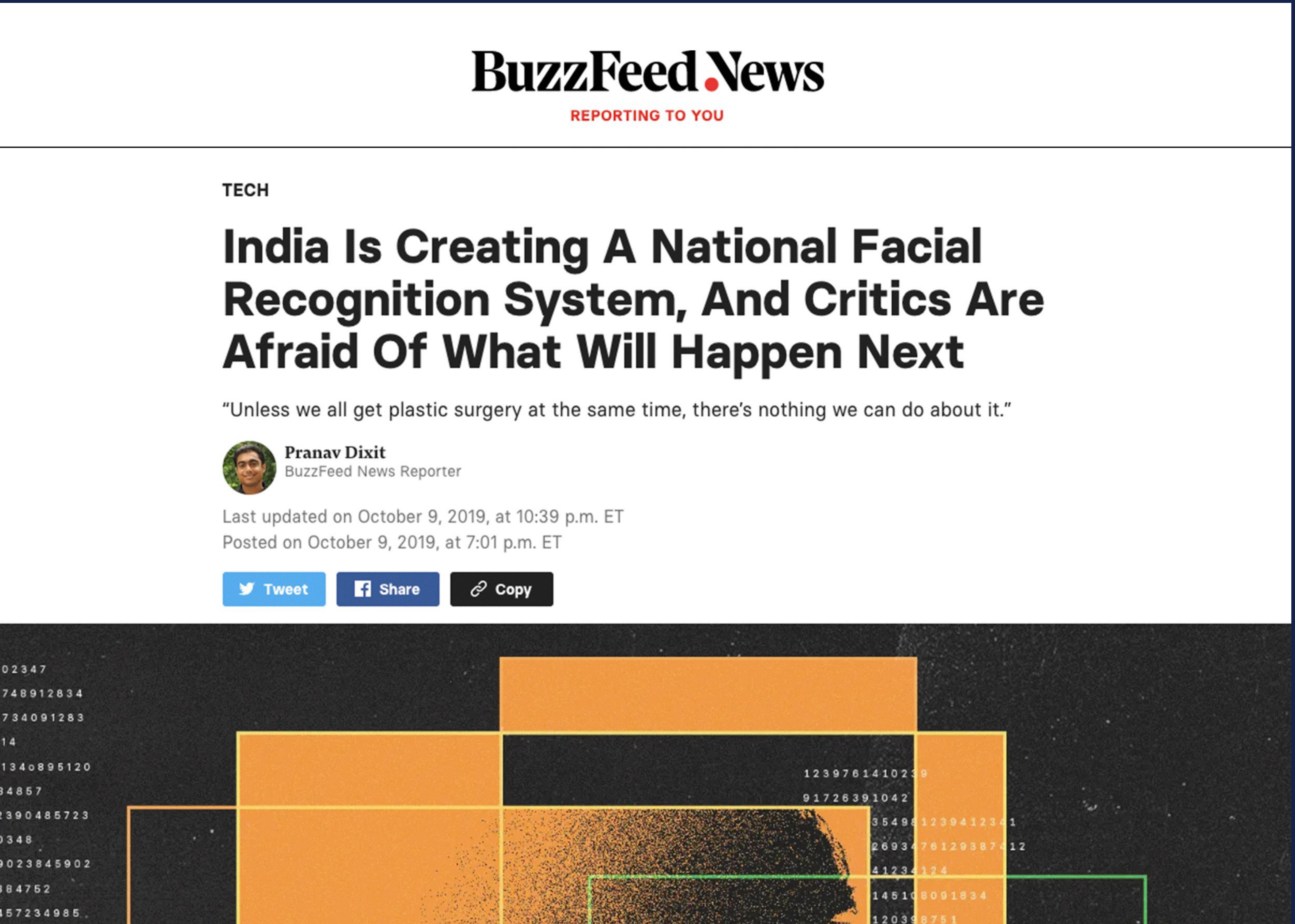
Pranav Dixit
BuzzFeed News Reporter

Last updated on October 9, 2019, at 10:39 p.m. ET
Posted on October 9, 2019, at 7:01 p.m. ET

Tweet Share Copy



Algorithmic Bias



BuzzFeed News
REPORTING TO YOU

TECH

India Is Creating A National Facial Recognition System, And Critics Are Afraid Of What Will Happen Next

"Unless we all get plastic surgery at the same time, there's nothing we can do about it."

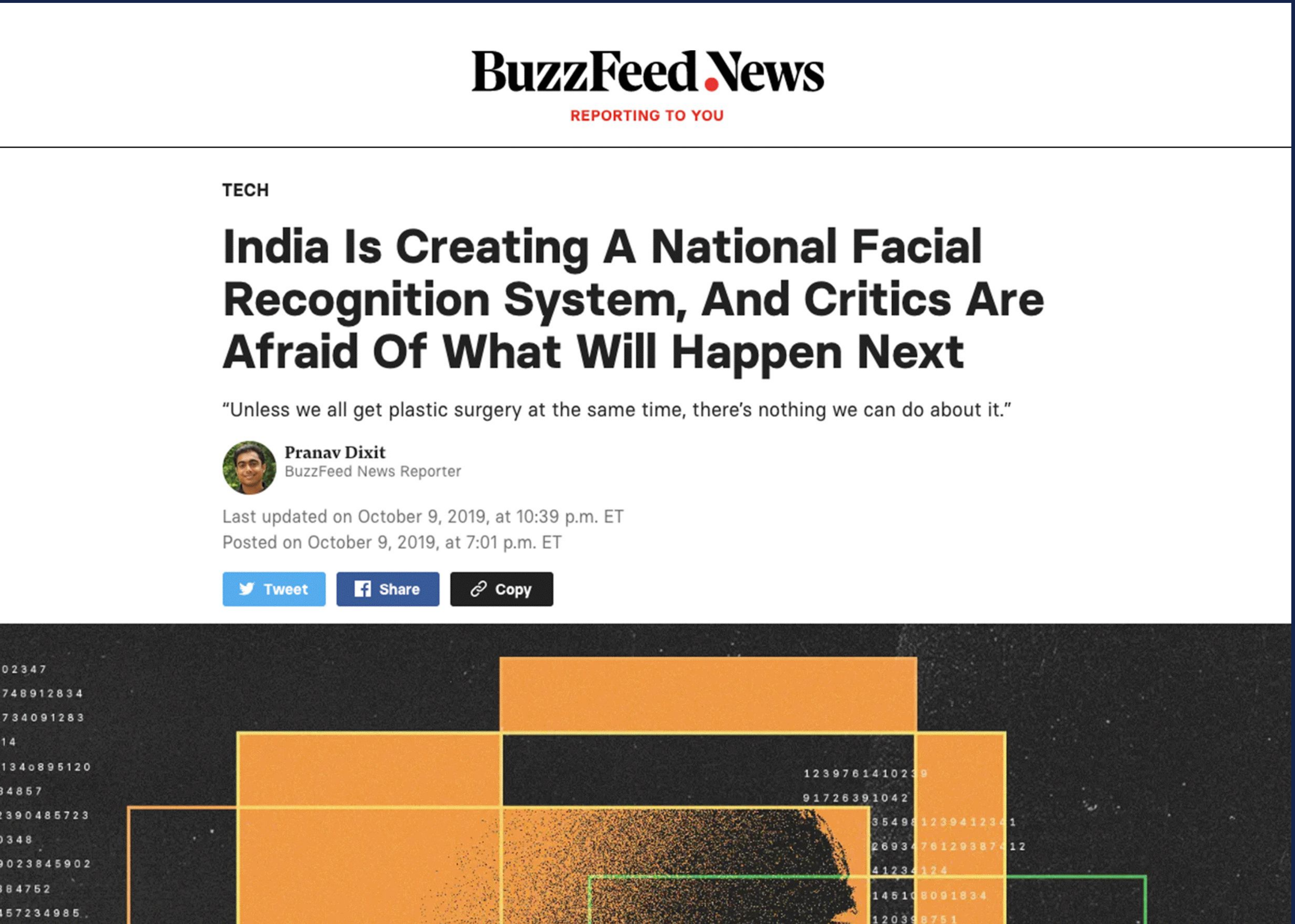
Pranav Dixit
BuzzFeed News Reporter

Last updated on October 9, 2019, at 10:39 p.m. ET
Posted on October 9, 2019, at 7:01 p.m. ET

Tweet Share Copy

Criminal Justice

Algorithmic Bias



BuzzFeed News

REPORTING TO YOU

TECH

India Is Creating A National Facial Recognition System, And Critics Are Afraid Of What Will Happen Next

"Unless we all get plastic surgery at the same time, there's nothing we can do about it."

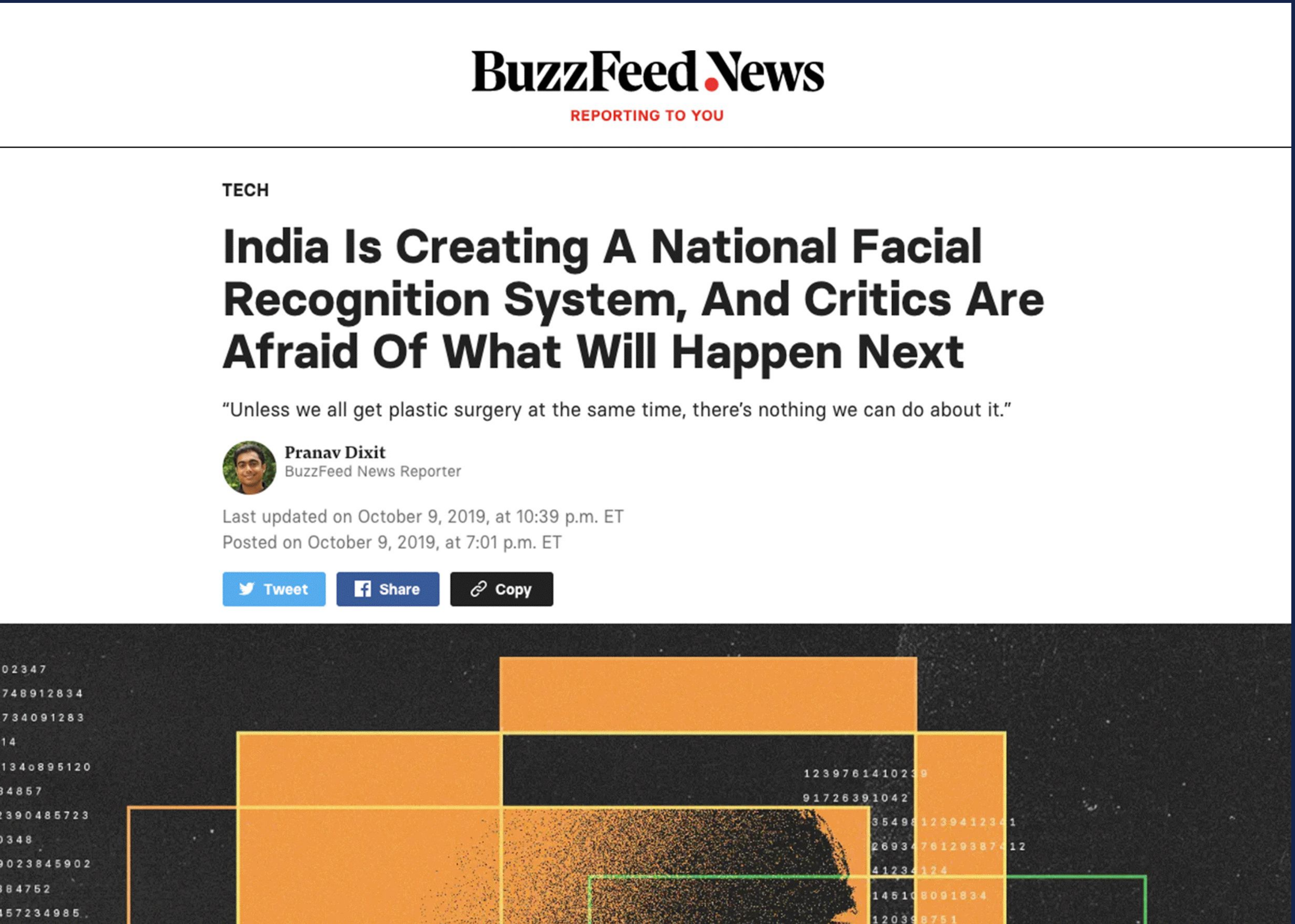
Pranav Dixit
BuzzFeed News Reporter

Last updated on October 9, 2019, at 10:39 p.m. ET
Posted on October 9, 2019, at 7:01 p.m. ET

Tweet Share Copy

- Criminal Justice
- Access to Jobs

Algorithmic Bias


BuzzFeed News

REPORTING TO YOU

TECH

India Is Creating A National Facial Recognition System, And Critics Are Afraid Of What Will Happen Next

"Unless we all get plastic surgery at the same time, there's nothing we can do about it."

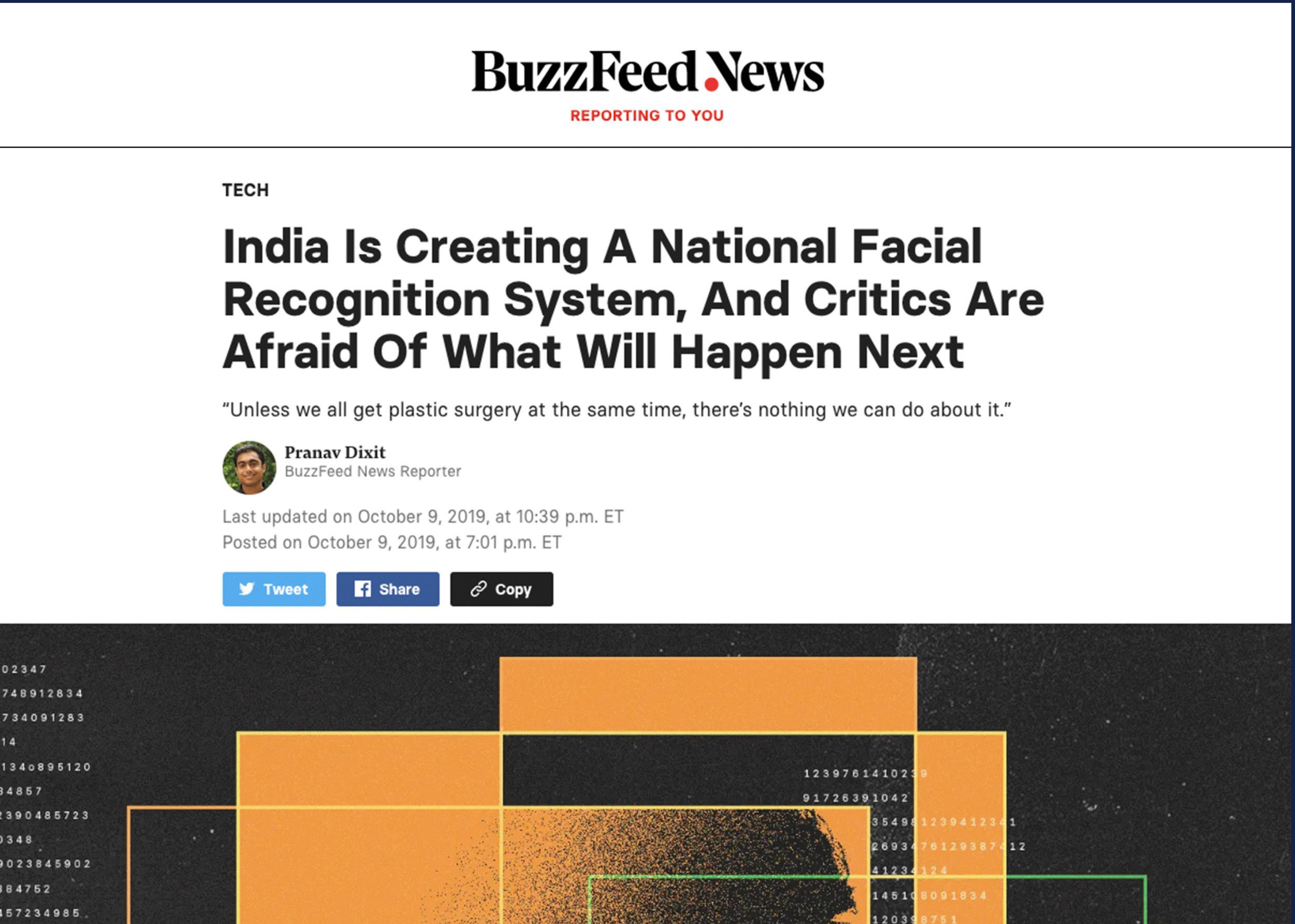
Pranav Dixit
BuzzFeed News Reporter

Last updated on October 9, 2019, at 10:39 p.m. ET
Posted on October 9, 2019, at 7:01 p.m. ET

Tweet | Share | Copy

- Criminal Justice
- Access to Jobs
- Image Recognition

Algorithmic Bias


BuzzFeed News

REPORTING TO YOU

TECH

India Is Creating A National Facial Recognition System, And Critics Are Afraid Of What Will Happen Next

"Unless we all get plastic surgery at the same time, there's nothing we can do about it."

Pranav Dixit
BuzzFeed News Reporter

Last updated on October 9, 2019, at 10:39 p.m. ET
Posted on October 9, 2019, at 7:01 p.m. ET

Tweet Share Copy

- Criminal Justice
- Access to Jobs
- Image Recognition
- Medical Diagnosis

Power and Responsibility

- Those who design and develop these technologies are in a position of power

Power and Responsibility

- Those who design and develop these technologies are in a position of power

- The choices about design and use which they make have a significant impact on the lives of others

Power and Responsibility

- Those who design and develop these technologies are in a position of power
- The choices about design and use which they make have a significant impact on the lives of others
- With this power comes responsibility

Science and Value

Science and Value

- Scientific research is not value-neutral

Science and Value

- Scientific research is not value-neutral
- It is a human activity governed by shared norms

Science and Value

- Scientific research is not value-neutral
- It is a human activity governed by shared norms
- These norms have profound effects

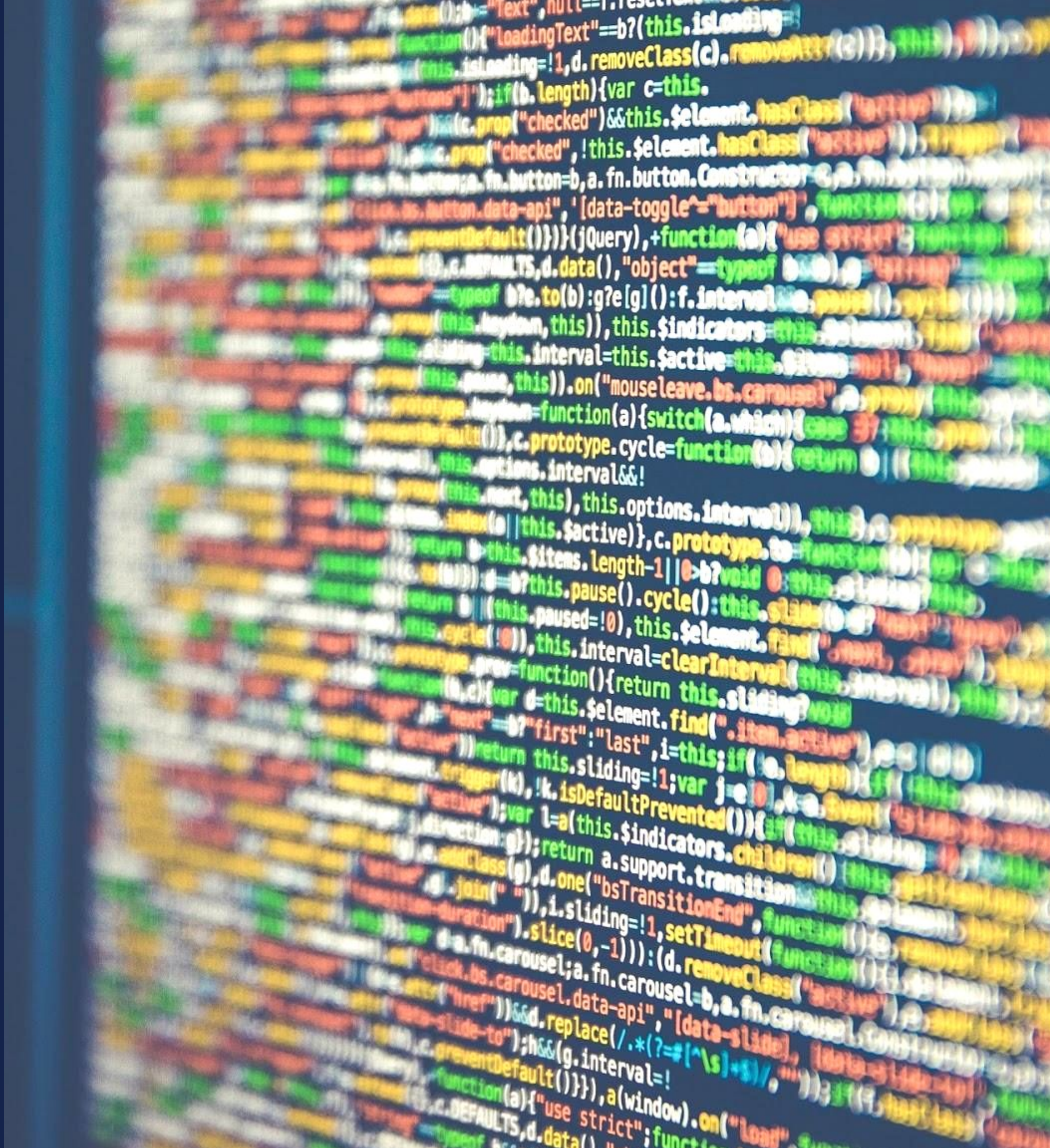
Responsible Innovation

Responsible Innovation

A transparent and iterative process by which societal actors and technologists become mutually responsive to each others needs to ensure the ethical and social value of scientific endeavour

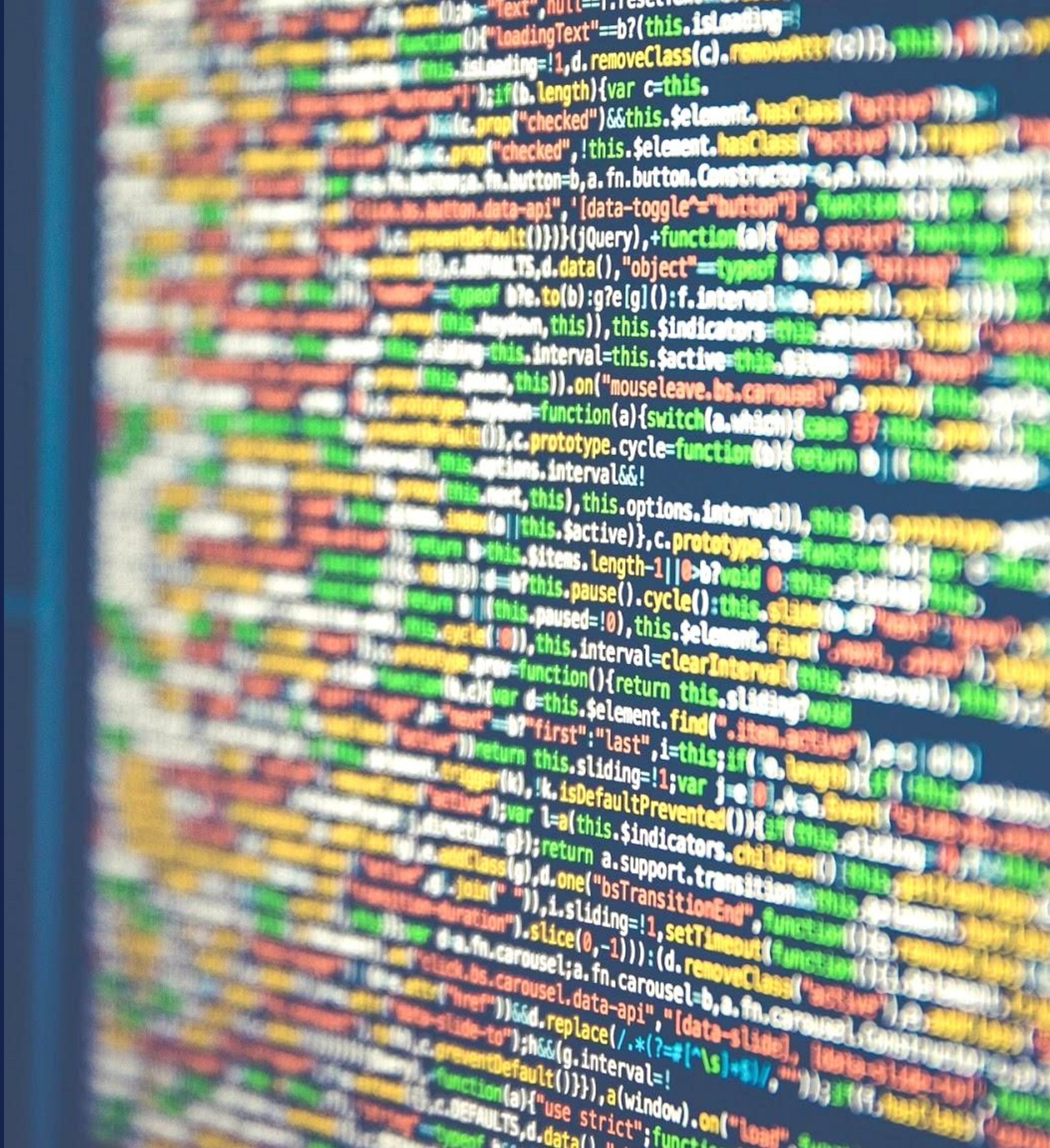
5 Principles and Process

The Responsibility of Technologists



The Responsibility of Technologists

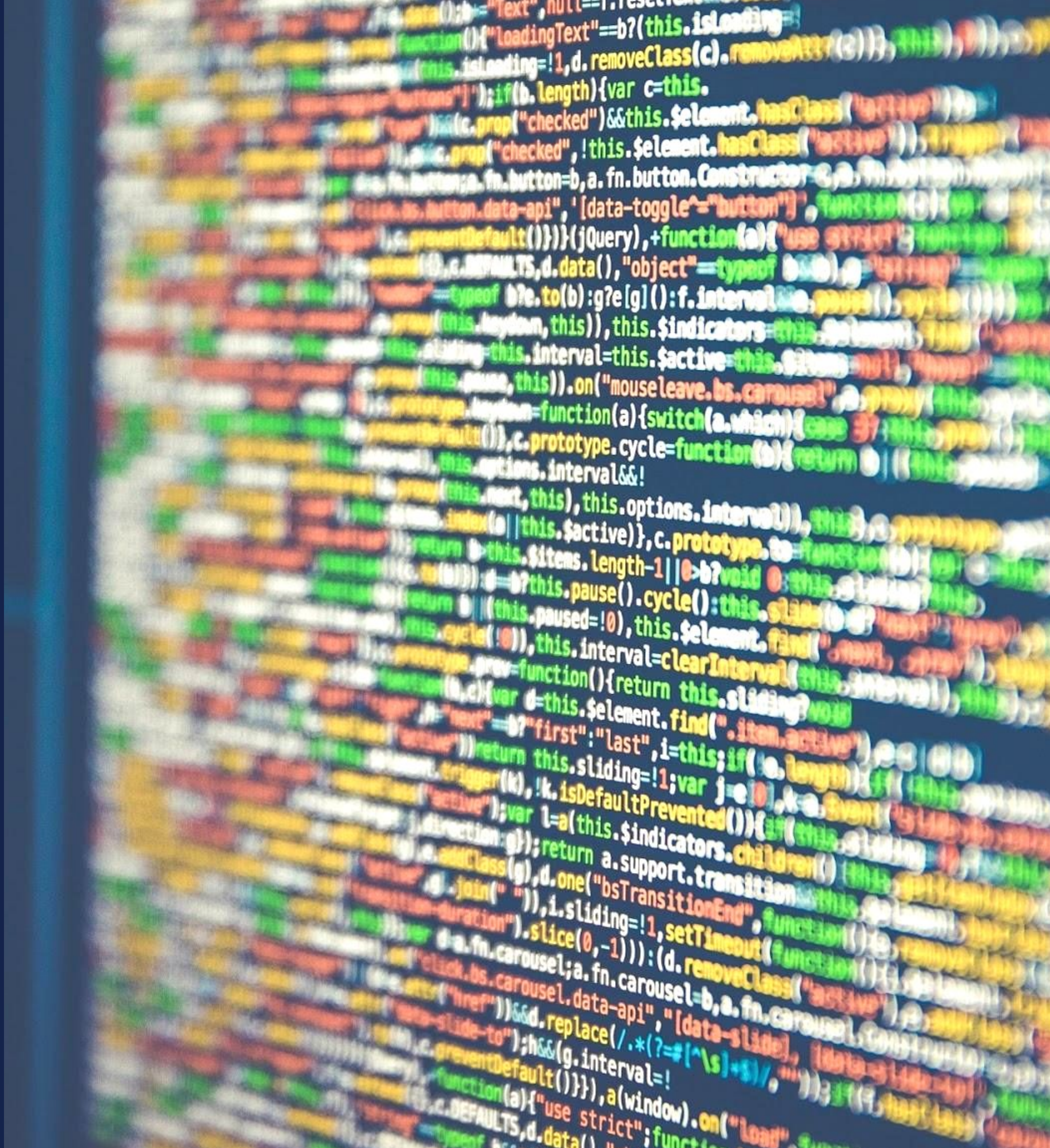
Intrinsic to the design of technology: technical safety, robustness, and ensuring that systems are accountable, transparent and fair from the start.



The Responsibility of Technologists

Intrinsic to the design of technology: technical safety, robustness, and ensuring that systems are accountable, transparent and fair from the start.

Extrinsic to the design of technology: intended use, mode of deployment in real-world settings, frameworks and institutions governing their application and use.



The AI Ethics Landscape

- 'Ethical Guidelines for trustworthy AI' – the European Union
- 'OECD Principles on AI' – OECD
- 'Beijing AI Principles' – Beijing Academy of Artificial Intelligence
- 'Asilomar AI Priniciples' – Future of Life Institute


Key Values

- Fairness, privacy and transparency

Credit: Kyle Glenn

Key Values

- Fairness, privacy and transparency
- Individual rights

Credit: Kyle Glenn

Key Values

- Fairness, privacy and transparency
- Individual rights
- Shared benefit from science


Credit: Kyle Glenn

Bridging the Gap

From principles to practice

A Five Step Process

The Five-Step Process

1. Does the technology have socially beneficial uses?

The Five-Step Process

1. Does the technology have socially beneficial uses?

2. Might this technology directly or indirectly result in harm?

The Five-Step Process

1. Does the technology have socially beneficial uses?

2. Might this technology directly or indirectly result in harm?

3. Is it possible to mitigate these risks, or eliminate them entirely?

The Five-Step Process

1. Does the technology have socially beneficial uses?
2. Might this technology directly or indirectly result in harm?
3. Is it possible to mitigate these risks, or eliminate them entirely?
4. With these measures in place, does the proposed action violate a 'red line' or moral constraint?

The Five-Step Process

1. Does the technology have socially beneficial uses?
2. Might this technology directly or indirectly result in harm?
3. Is it possible to mitigate these risks, or eliminate them entirely?
4. With these measures in place, does the proposed action violate a 'red line' or moral constraint?
5. With these measures in place, do the benefits outweigh the risks?

1. Does the technology have socially beneficial uses?

1. Does the technology have socially beneficial uses?

- Well-being
- Autonomy
- Justice
- Public institutions
- Global challenges

2. Might this technology directly or indirectly result in harm?

2. Might this technology directly or indirectly result in harm?

- Undermine health, well-being or human dignity
- Restrict freedom or autonomy
- Lead to unfair treatment or outcomes
- Harm public institutions or civic culture
- Infringe human rights

3. Is it possible to mitigate these risks or eliminate them entirely?

3. Is it possible to mitigate these risks or eliminate them entirely?

Control the release of technologies or the flow of information

3. Is it possible to mitigate these risks or eliminate them entirely?

- Control the release of technologies or the flow of information
- Adopt technical solutions and countermeasures

3. Is it possible to mitigate these risks or eliminate them entirely?

- Control the release of technologies or the flow of information
- Adopt technical solutions and countermeasures
- Help the public understand new technologies

3. Is it possible to mitigate these risks or eliminate them entirely?

- Control the release of technologies or the flow of information
- Adopt technical solutions and countermeasures
- Help the public understand new technologies
- Seek out policy solutions and legal frameworks

4. With these measures in place, does the proposed action violate a 'red line' or moral constraint?

- Consent
- Weapons
- Surveillance
- International law and human rights

4. With these measures in place, does the proposed action violate a 'red line' or moral constraint?

5. With these measures in place, do the benefits outweigh the risks?

Two final tests

Two final tests

- Have you thought about everyone who could be affected by your action?

Two final tests

- Have you thought about everyone who could be affected by your action?
- Might you have reason to regret it later?

5 The Path Ahead

Key Ideas

Key Ideas

- Those who design and develop these technologies have a responsibility to think about how they will be used.

Key Ideas

- Those who design and develop these technologies have a responsibility to think about how they will be used.

- There are concrete steps and processes that we can put in place to make sure this responsibility is successfully discharged.

Key Ideas

- Those who design and develop these technologies have a responsibility to think about how they will be used.
- There are concrete steps and processes that we can put in place to make sure this responsibility is successfully discharged.
- We are responsible for what we can reasonably foresee and should take steps to bring about positive outcomes

New Directions

Research

- Research
- Norms

- Research
- Norms
- Practice

Thank you

Technology can "lock in" value

Designing for social effect

Robert Moses, New York City (1930s)

Technology can “lock in” value

Designing for social effect

Robert Moses, New York City (1930s)

Ethics and Machine Learning

Ethics and Technoloy