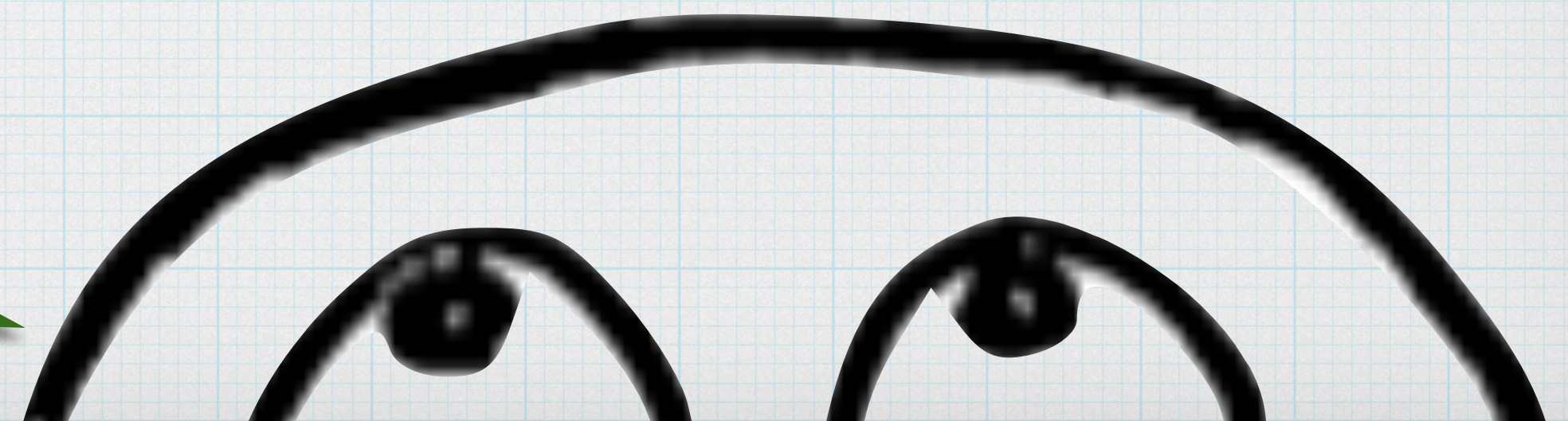


★ ex15 の B

機械学習・AIと社会

機械学習I/II

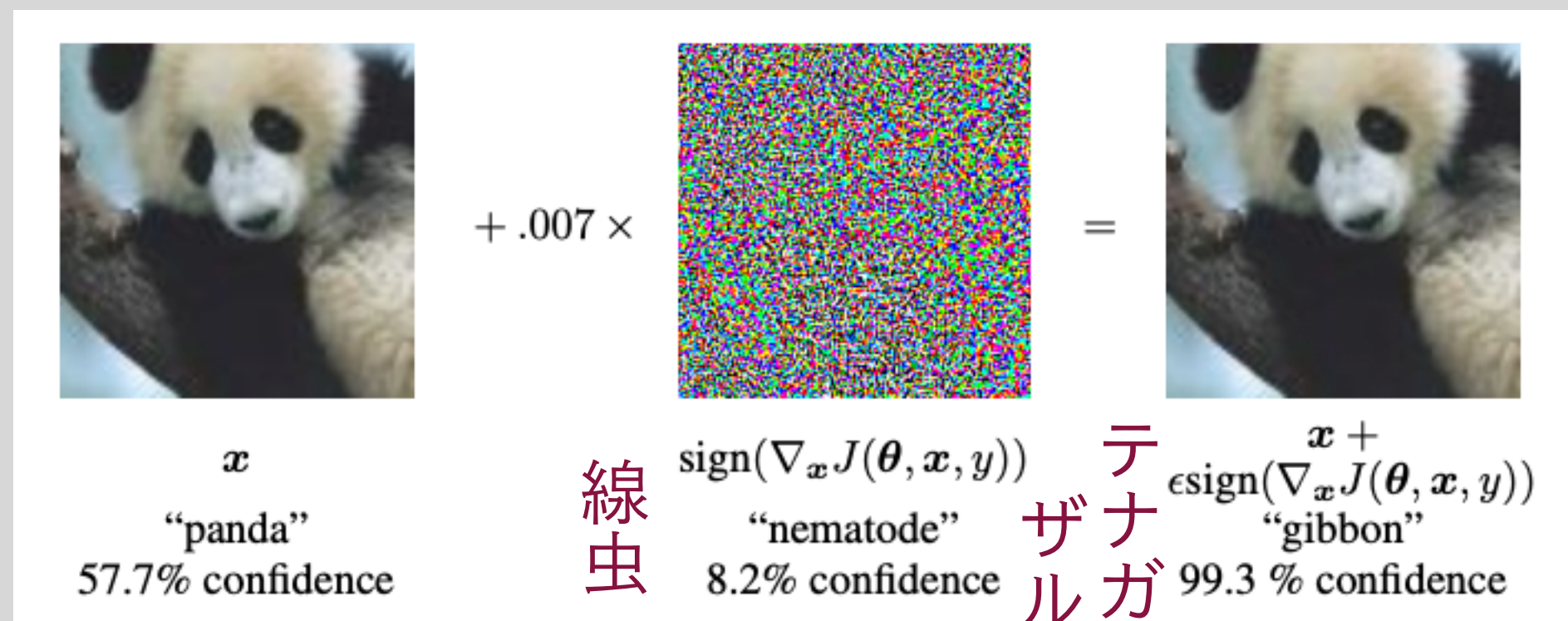




機械学習・AIシステムの信頼性を高めるには，出力の再現性・ロバスト性，説明可能性が重要

## 再現性・ロバスト性

- 再現性: 一定の条件下でシステムが一貫した出力を返せる（同じ入力に対して同じ出力をする）か
- ロバスト性: 異常な状況（異常な条件，異常なデータ等）に対して頑健（ロバスト）か



I.J. Goodfellow, et al. "Explaining and Harnessing Adversarial Examples," ICPR 2015 より引用

画像認識を学習した畳み込みNNを  
だます「敵対的事例」を作る研究

ロバスト性は，例外値やエラーへの対応  
だけでなく，意図的な攻撃に対する防御  
の面でも重要

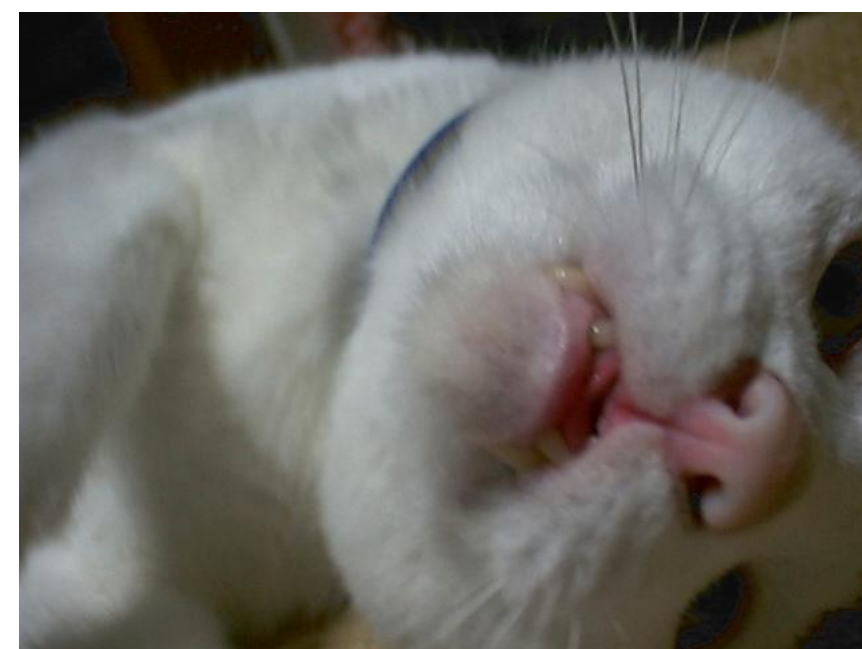


機械学習・AIシステムの信頼性を高めるには，出力の再現性・ロバスト性，説明可能性が重要

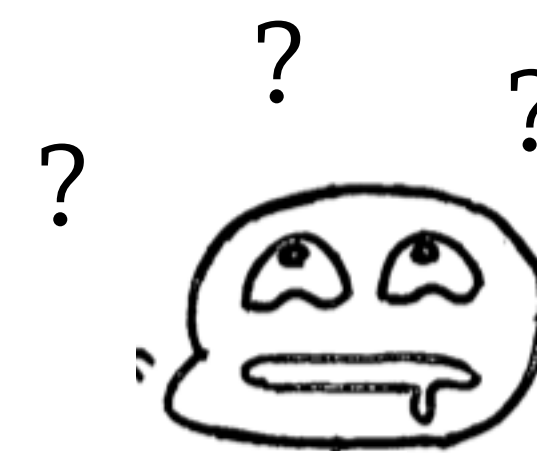
### 説明可能性

システムが生み出す出力について，なぜそのような出力をしたのかを人間のユーザに説明できる能力（人間が解釈・理解しやすいかどうか）．

現在主流の機械学習のアルゴリズム，特に大規模な深層ニューラルネットは，パラメータ数が多く複雑な構造をしており，解釈・理解が難しい．



"ねこ"



出力は正しいけれど  
どうしてかは謎。  
どこをどう判断？



### （よだんだよん） chat AI は "hallucination"（幻覚）をみる

（単純化した説明なので不正確だが...）大規模言語モデルは、大量の文章を虫食いにしたものを入力として、隠された部分に入る確率の高い言葉を学習している。chat AI は、それをもとにして、与えられた文章に続く確率の高い（それっぽい）文章を生成している。文法の正確さや文章の自然さは高いが、内容の正確さが保証されるものではない。

ユーザの質問: "フリードリヒ・ニーチェは彼の哲学をどのように宇宙旅行に適用しましたか？"

AIの応答: "フリードリヒ・ニーチェは、宇宙旅行の概念を彼の存在論や道徳哲学に組み込みました。彼は、人間が新たな環境で生き抜くためには自我を超越する必要があると主張しました。これは彼の「超人」の理念につながっており、彼にとって宇宙旅行はその究極の表現形態でした。"

しかし、実際には、フリードリヒ・ニーチェは宇宙旅行の概念を彼の哲学に組み込んでいません。ニーチェが活動していた19世紀には宇宙旅行はまだ現実のものではなかったため、彼の著作にそのようなテーマが存在するはずがありません。これは完全なハルシネーションで、AIが虚偽の情報を生成してしまっています。

実は、この文章全体が、chatGPT(GPT4)に「ハルシネーションの例を作って」といった結果...



- ・ 機械学習・AIのシステムが偏った（バイアスのある）出力をすることがある
- ・ システムの公平性を測定する／偏りの少ないシステムを作るための技術や，法的規制の整備が必要

### 問題となった事例

#### 某社の画像認識アプリ

特定の肌の色の人をゴリラと識別する事例が発生 ⇒ 提供中止

原因は，ヒトの顔画像の学習データが白人に偏っており，肌の色によって識別精度に大きな違いが出たこと

#### 某社のAI人事システム

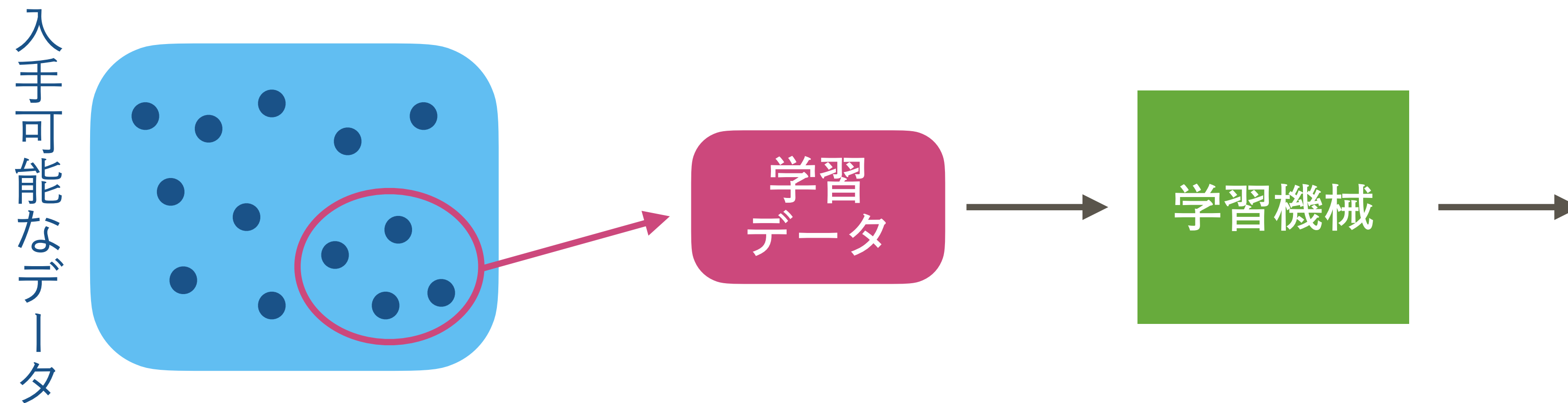
エンジニアの採用人事を効率化するために，履歴書から評価スコアを予測するシステムを実用化．女性というだけで評価スコアが低くなることが判明し，使用中止に

原因は，学習データが極端に男性のものに偏っていたこと



## 機械学習・AIの公平性 2/2

- ・ 機械学習・AIのシステムが偏った（バイアスのある）出力をすることがある
- ・ システムの公平性を測定する／偏りの少ないシステムを作るための技術や，法的規制の整備が必要



- ・ 学習データが偏ったサンプルになっていないか？
- ・ 入手可能なデータがそもそも偏っていないか？
- ・ （データの偏りとは独立に）出力が公平性を持っているか？



「ビッグデータ」の回にはデータそのものの扱いについて考えたが、データを学習に利用した結果生ずる問題もあることに注意

### プライバシーの問題

- システム作成者の立場では、同意を得て情報収集する必要があるかもしれないこと、収集した情報が漏洩しないようにすること、ユーザが特定されてしまわないよう匿名化等の処置を施す必要があること、等に要注意.
- ユーザは、適切なプライバシー保護を行わない相手に情報を提供しないよう防衛する必要がある. また、むやみな情報収集や悪用を防ぐ法規制も必要となる.

### セキュリティの問題

- データの保護: (システム側) データを漏洩させない / (ユーザ側) 機密情報をデータとして提供しない
- 攻撃に対する防衛: 一般的なコンピュータシステムのセキュリティ問題に加えて、学習過程やシステムの出力を歪めようとする攻撃にも注意が必要 (cf. adversarial examples)

### 悪用

例: 機械学習・AIの技術を用いて、特定人物の (存在しない) 映像や音声を作り出すことが可能.  
詐欺や情報操作に悪用される



「ビッグデータ」の回にはデータそのものの扱いについて考えたが、データを学習に利用した結果生ずる問題もあることに注意

（「安全性」とはちょっと視点が違うけど...）

### 著作権に関する問題

- 機械学習・AIシステムが画像や楽曲等のコンテンツを生成した場合、その著作権は誰に帰属するのか
- 学習データの中に著作権で保護されるものがあった場合、学習したシステムの出力物が著作権を侵害しているとみなされる可能性がある（場合によってはデータとして学習に利用することそのものも。 国によって法律が異なることに注意）

文化庁と内閣府が 2023-05-30 に公開した「AIと著作権の関係等について」

[https://www8.cao.go.jp/cstp/ai/ai\\_team/3kai/shiryo.pdf](https://www8.cao.go.jp/cstp/ai/ai_team/3kai/shiryo.pdf)

- AIと著作権の関係は、「生成・利用段階」と「AI開発・学習段階」を分けて考えるべき
- AIによって生成したコンテンツを公開したり販売したりする場合、通常の著作物と同様に扱われる（既存の著作物との類似性や依拠性があれば著作権侵害と認められ得る）