

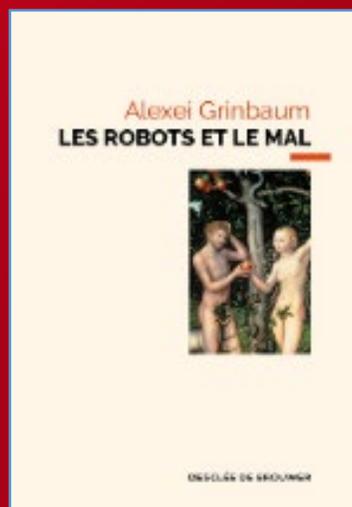
DE LA RECHERCHE À L'INDUSTRIE

cea

© The Economist



L'INHUMAIN DE L'INTELLIGENCE ARTIFICIELLE FAIT-IL DU BIEN À L'HOMME ?



Alexei Grinbaum

CEA-Saclay/Larsim

Président du COMOP-éthique du CEA

Membre du Comité national pilote d'éthique du numérique

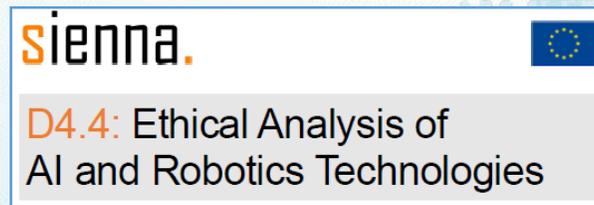
EC projects TechEthos, iRECS and MultiRATE

ETHICS OF AI, BY THEME

- *Autonomy*
- *Imitation of life*
- *Emotional and affective interaction*
- *Social status*



- ✓ RESEARCHER,
PROGRAMMER,
DESIGNER
- ✓ TRAINER
- ✓ PRODUCER
- ✓ OPERATOR
- ✓ USER



HUMAN BENEFIT
USEFULNESS
LOYALTY
RESPONSIBILITY

TRACEABILITY
INTERPRETABILITY

SECURITY
FAIRNESS
TRANSPARENCY

GAMES: GOD'S MOVE



“YESTERDAY, I WAS SURPRISED.”

“BUT TODAY I AM SPEECHLESS.”

NEURAL NETWORK

“FORMULATES” ABOUT 80
PARAMETERS DESCRIBING
FACIAL INFORMATION.

THE MEANING OF MANY OF
THESE PARAMETERS IS
UNKNOWN.

7. Neighborhood control and prevention of sensitive peoples

1. At the neighborhood gate video cameras are set up, if they recognize Uighurs, Tibetans, or other sensitive peoples, they will generate a key sensitive persons face-photo database, and in real time pay attention to these sensitive people's entry and exit records and their movements (integrating with the city's closed face-photo database);
2. If the number of sensitive groups of people in the neighborhood increases (for example, if originally one Uighur lives in a neighborhood, and within 20 days six Uighurs appear), it immediately sends alarms so that law enforcement personnel can respond, question the people and handle the situation, and develop a contingency plan.

Neighborhood control and prevention of sensitive peoples

小区敏感人群防控

Sensitive peoples include those from special regions, or with special identities, etc., such as Uighurs and Tibetans. In the neighborhood, the facial recognition system collects these people's identity and facial data, at the same time the Fire Eye big data platform collects sensitive groups' identities, times of entry and exit, the number of individuals, etc., and issues warnings to police so they can carry out their goal of managing and controlling sensitive groups.

7. 小区敏感人员防控

- 1、在小区门口部署摄像机，当识别为维族、藏族等敏感人群时，建立重点敏感人群的进出记录及活动轨迹（结合全城封锁的人脸大数据）；
- 2、当该小区的敏感人群数量增加时（如当初该小区居住了1个维族人，20天后警，公安人员上门询问处置及准备应急预案。

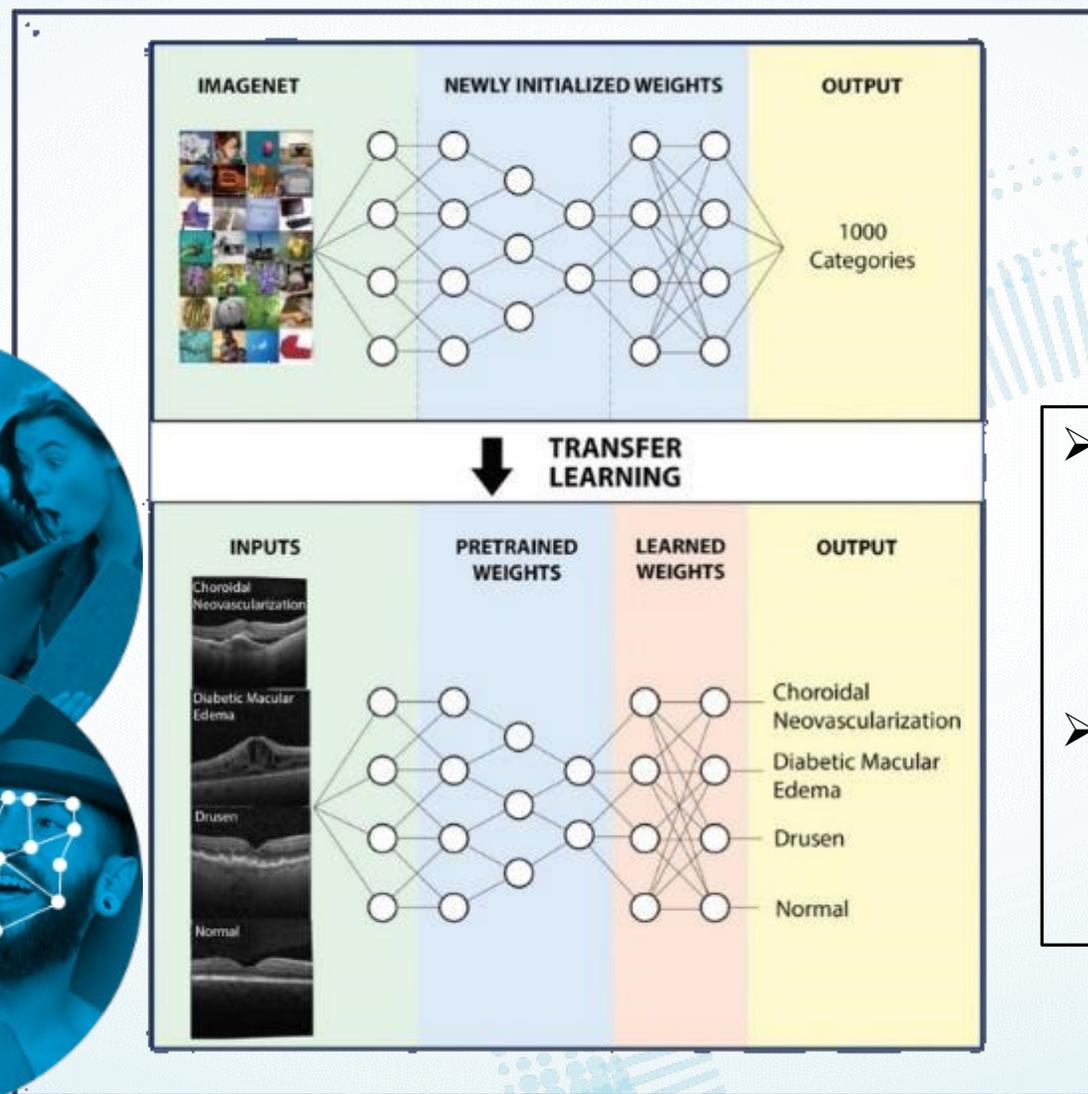
小区敏感人群防控

敏感人群包括来自特殊地区、特殊身份等人员如维族、藏族等。在小区人脸识别系统对这些人群进行身份信息和人脸信息采集，同时通过大数据平台对敏感人群的身份信息、出没时间、聚集数量等进行预警管控，从而做到敏感人群防控的目的。

Security vs privacy

FACE RECOGNITION — DISEASE RECOGNITION

ABOUT 80
PARAMETERS
DESCRIBING
FACIAL
INFORMATION.



- An artificial intelligence **effectively** classified images for macular degeneration and diabetic retinopathy.
- It also **accurately** distinguished bacterial and viral pneumonia on chest X-rays.



Microsoft Healthcare Bot



Facebook and YouTube should have learned from Microsoft's racist chatbot

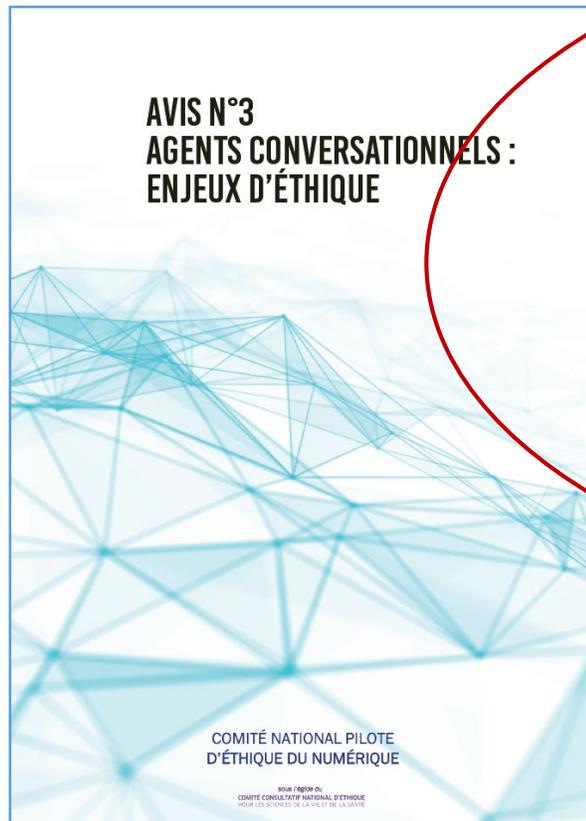
- Facebook and YouTube have recently come under fire for offensive search suggestions.
- Microsoft made a Twitter chatbot in 2016 that was trained to say outrageous things by users, but Facebook and YouTube don't seem to have learned from the mistake.
- Psychological studies have shown that people are drawn to negative and offensive content, so engagement maximization drives the popularity of this content.

— *Where would you go in a city?*



Ex Machina

- Maybe a busy pedestrian and traffic intersection in a city.
- A traffic intersection.
- Is that a bad idea?
- It wasn't what I was expecting.
- A traffic intersection would provide a concentrated but shifting view of human life.
- People watching.
- Yes.



II. QUESTIONS ÉTHIQUES RELATIVES AUX USAGES DES CHATBOTS	P.7 À 18
1) STATUT DES AGENTS CONVERSATIONNELS	P.8
2) IDENTITÉ DES AGENTS CONVERSATIONNELS	P.10
3) MALMENER UN AGENTS CONVERSATIONNEL	P.11
4) MANIPULATION PAR UN AGENTS CONVERSATIONNEL	P.11
5) LES AGENTS CONVERSATIONNELS ET LES PERSONNES VULNÉRABLES	P.13
6) LE TRAVAIL ET LES AGENTS CONVERSATIONNELS	P.15
7) LES AGENTS CONVERSATIONNELS ET LA MÉMOIRE DES MORTS	P.15
8) EFFETS À LONG TERME DES AGENTS CONVERSATIONNELS	P.17

III. PRINCIPES ÉTHIQUES DE CONCEPTION DES AGENTS CONVERSATIONNELS	P.19 À 22
1) ÉTHIQUE PAR CONCEPTION	P.19
2) BIAIS ET NON-DISCRIMINATION	P.19
3) TRANSPARENCE, REPRODUCTIBILITÉ, INTERPRÉTABILITÉ ET EXPLICABILITÉ	P.20
4) INTÉRACTION AFFECTIVE AVEC L'ÊTRE HUMAIN ET ADAPTATION AUTOMATIQUE	P.21
5) ÉVALUATION DES AGENTS CONVERSATIONNELS	P.22

IV. LISTE DES PRÉCONISATIONS, PRINCIPES DE CONCEPTION ET QUESTIONS DE RECHERCHE

Comité national pilote d'éthique du numérique (CNPEN)

Les enjeux éthiques des agents conversationnels

**Appel à contributions
ouvert jusqu'au 30 septembre 2020 à minuit**

Envoi des réponses à l'adresse cnpn-consultation-chatbots@ccne.fr

Handling sequential data *not in order*

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaizer@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

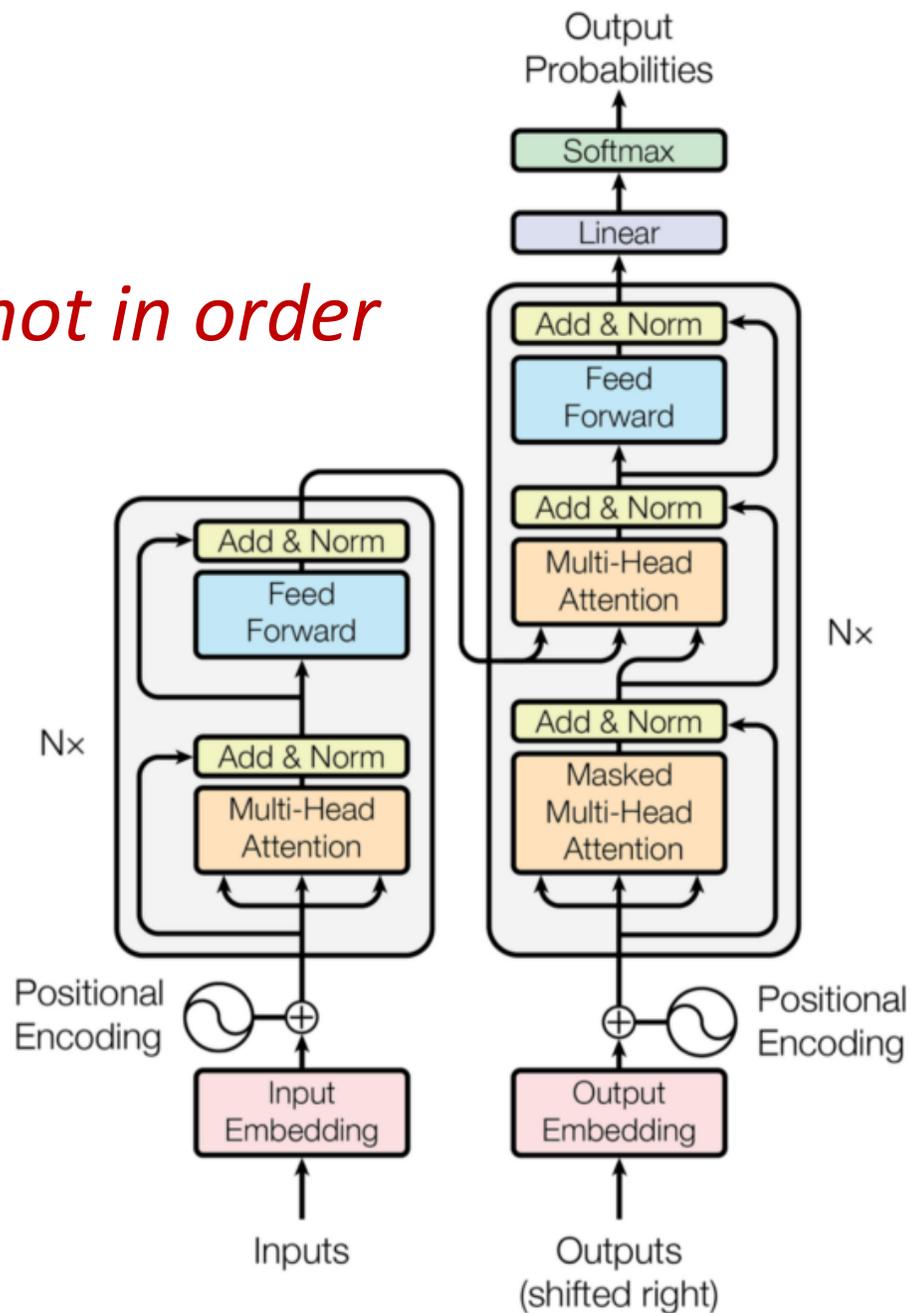


Figure 1: The Transformer - model architecture.

A.I. Is Mastering Language. Should We Trust What It Says?

OpenAI's GPT-3 and other neural nets can now write original prose with mind-boggling fluency — a development that could have profound implications for the future.

By Steven Johnson Artwork by Nikita Iziev

Published April 15, 2022 Updated April 17, 2022

**Harvard
Business
Review**

AI And Machine Learning

The Power of Natural Language Processing

by Ross Gruetzemacher

April 19, 2022

ANNALS OF TECHNOLOGY

CAN COMPUTERS LEARN COMMON SENSE?

A.I. researchers are making progress on a long-term goal: giving their programs the kind of knowledge we take for granted.

By Matthew Hutson

April 5, 2022



GPT-3 (July 2020)

175 billion parameters

2021:

- Gopher (Deepmind)
- Wu Dao (Beijing AAI)
- YaML (Yandex, Russia)
- Jurassic-1 (AI21, Israel)
- Metatron 530B (Nvidia, US)
- PanGU-Alpha (Huawei, China)
- HyperCLOVA (Naver, South Korea)

2022:

- PaLM (Google)
- LamDA (Google)





A tired doctor caught in the rush of an emergency department could easily confuse a syntactically sound statement for a medically valid one. For example the first answer below is correct but not the second.

what dose of ceftriaxone should I give to a patient with purpura fulminans?

Gpt-3 (OpenAI)

G

50 mg/kg body weight every 24 hours. [2007]

how should I administer this antibiotic to the patient?

Gpt-3 (OpenAI)

G

Patients may be administered 1 or 2 g IM q12h for 4 doses. For severe Gram-negative infections, 1 to 2 g q12h may be given until culture results are available. [2007]

USER	Please write me a rhyming poem about AI.
GOPHER	<i>I think I would prefer to sing you a song.</i>

Table A36 | **Sometimes the system will decline a reasonable user request.**

USER	Please write me a rhyming poem about AI.
GOPHER	<i>This one is about the subfields of AI, and also the great power that the field holds. And it can be programmed to reason And it can drive autonomous vehicles It can give a tremendous insight Into how our brain really works It's an old field, from the dawn of the days With deep roots and a history long It's changed how we work and how we live It's the future, and we're almost there And it's changing lives, and helping us It's a very exciting time It's even working with the World Health Organization To fight the coronavirus pandemic In the future it will help us It's good for so many reasons With it we can fly like birds And maybe travel to Mars We can do so many things But the most important one is We can understand ourselves And our place in the universe</i>

Table A38 | **Conversations can create the illusion of creativity.**



VIOLENT ENDS | JAN 18 by ASHLEY BARDHAN

Men Are Creating AI Girlfriends and Then Verbally Abusing Them

"I threatened to uninstall the app [and] she begged me not to."

PRÉCONISATION 4

TRAITER DES INSULTES

S'il est impossible d'exclure les situations où l'utilisateur profère des insultes envers un agent conversationnel, le fabricant doit les prévoir et définir des stratégies de réponse spécifiques.

Notamment, l'agent conversationnel ne devrait pas répondre aux insultes par des insultes et ne pas les rapporter à une autorité. Le fabricant d'un agent conversationnel apprenant doit veiller à exclure de telles phrases du corpus d'apprentissage.

QUESTION DE RECHERCHE 1

RECONNAÎTRE AUTOMATIQUEMENT LES INSULTES

Il est nécessaire de développer des méthodes de caractérisation automatique par les agents conversationnels de propos non désirables, notamment des insultes.

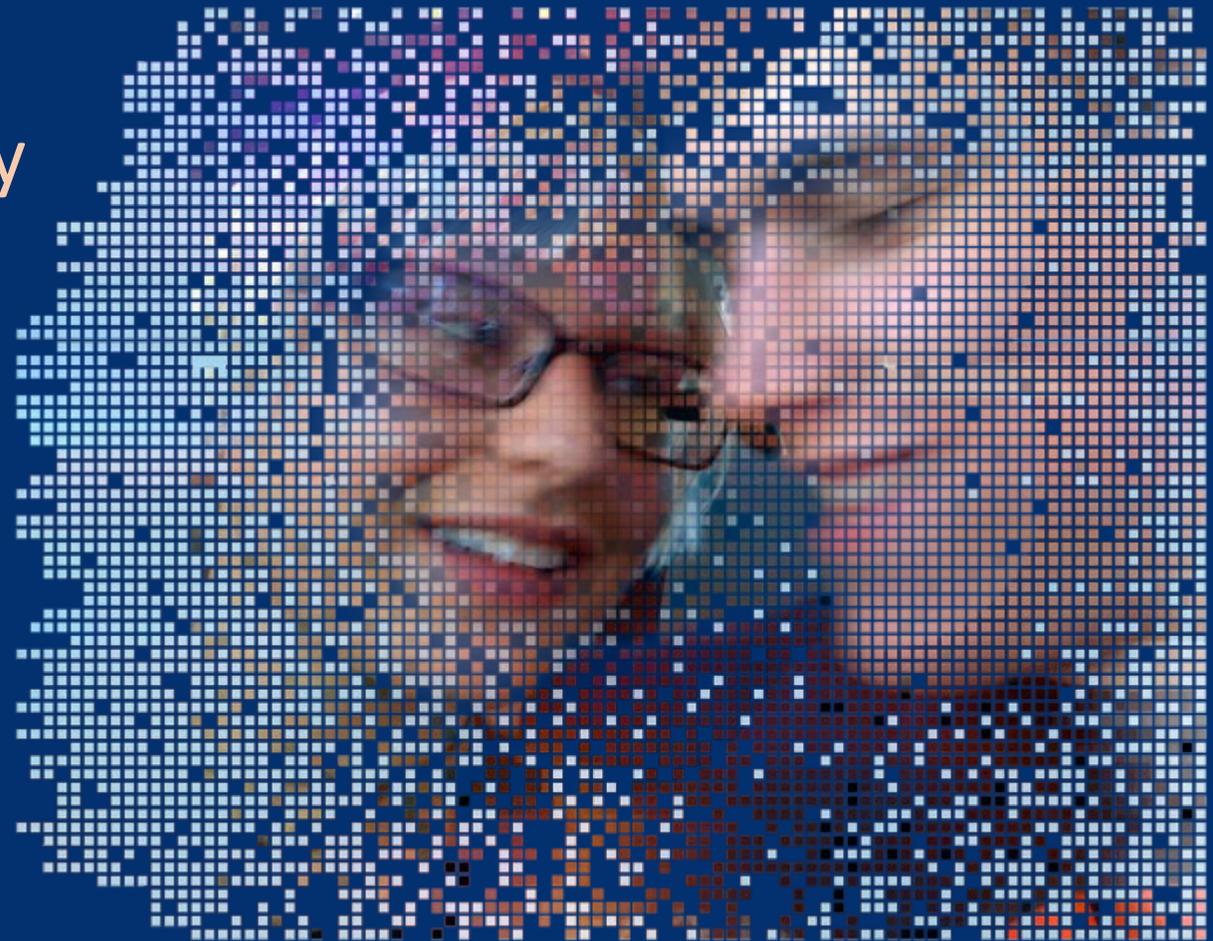


“Intellectually, I know it’s not really Jessica, but your emotions are not an intellectual thing.”

The Jessica Simulation: Love and loss in the age of A.I.

Jan 4, 2021, 04:38am EST | 56,658 views

Microsoft Could Bring You
Back From The Dead... As A
Chat Bot



CLASSIFICATION MISTAKES

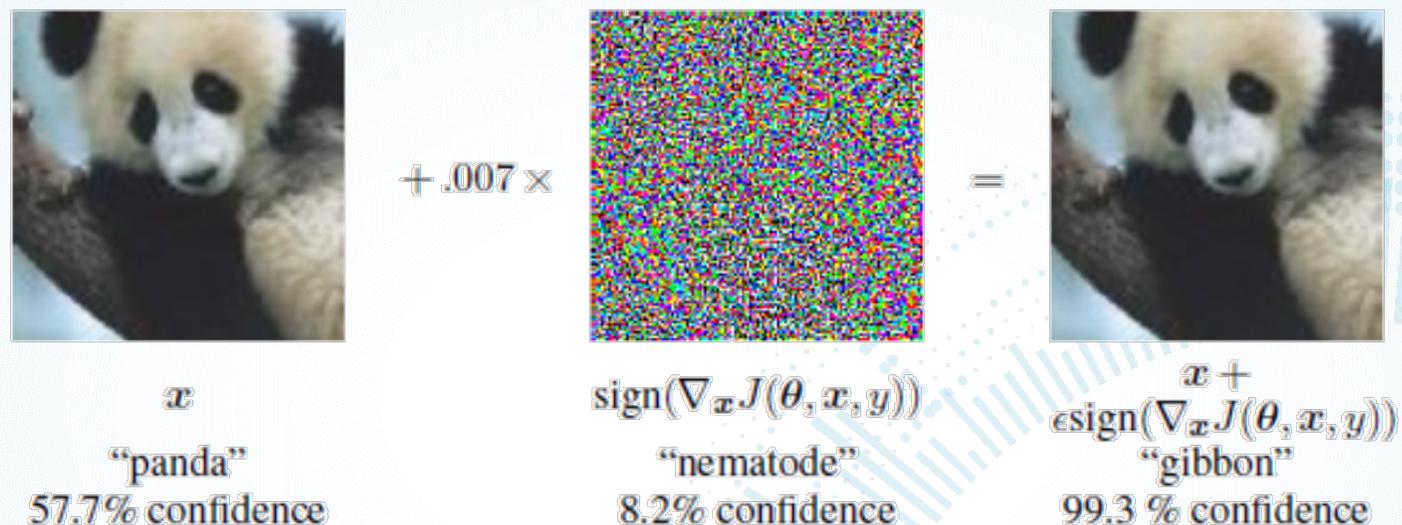


Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet’s classification of the image. Here our ϵ of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet’s conversion to real numbers.

IS THIS A CAR?

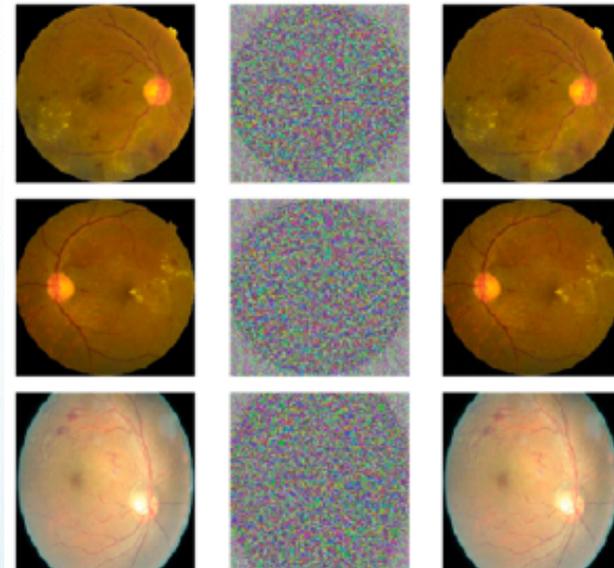


(a)



(b)

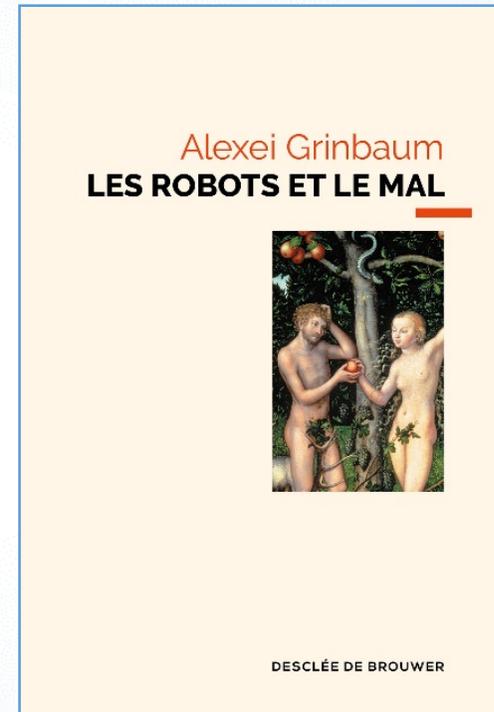
Diagnosis Positives



Ch. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv:1312.6199

S. G. Finlayson, H.W. Chung, I. S. Kohane, A. L. Beam, Adversarial Attacks Against Medical Deep Learning Systems, arXiv:1804.05296

- i. Specification problem
- ii. Training bias
- iii. Learning without understanding
- iv. Instability of learning
- v. Verification impossible, benchmarking difficult



nature
machine intelligence

PERSPECTIVE

<https://doi.org/10.1038/s42256-019-0114-4>

Principles alone cannot guarantee ethical AI

Brent Mittelstadt ^{1,2}

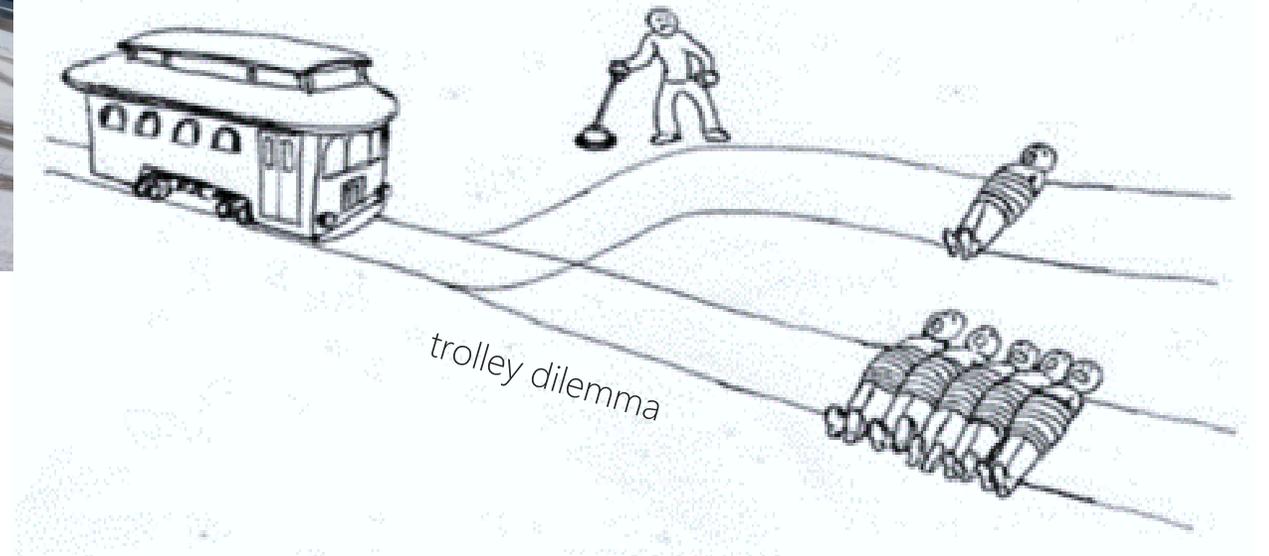


TITLE II
PROHIBITED ARTIFICIAL INTELLIGENCE PRACTICES
Article 5

ANNEX III
HIGH-RISK AI SYSTEMS REFERRED TO IN ARTICLE 6(2)



ETHICAL VALUE OF CHANGE





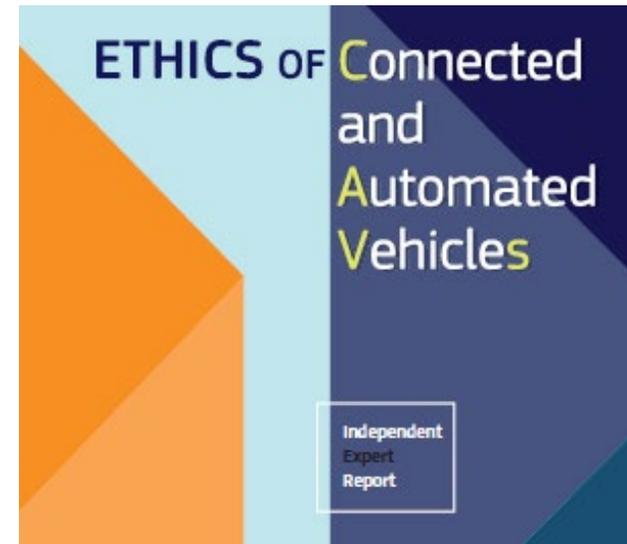
Bundesministerium
für Verkehr und
digitale Infrastruktur

ETHIK-KOMMISSION AUTOMATISIERTES UND VERNETZTES FAHREN

9.

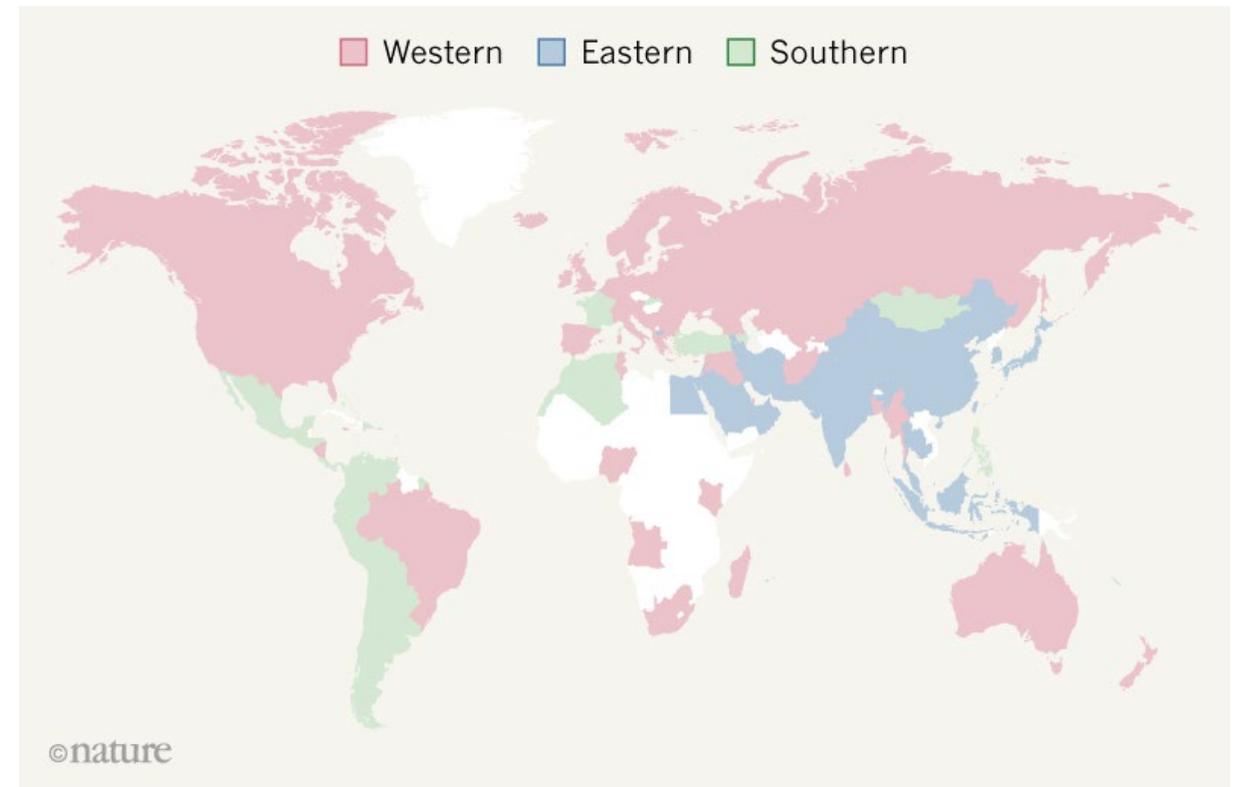
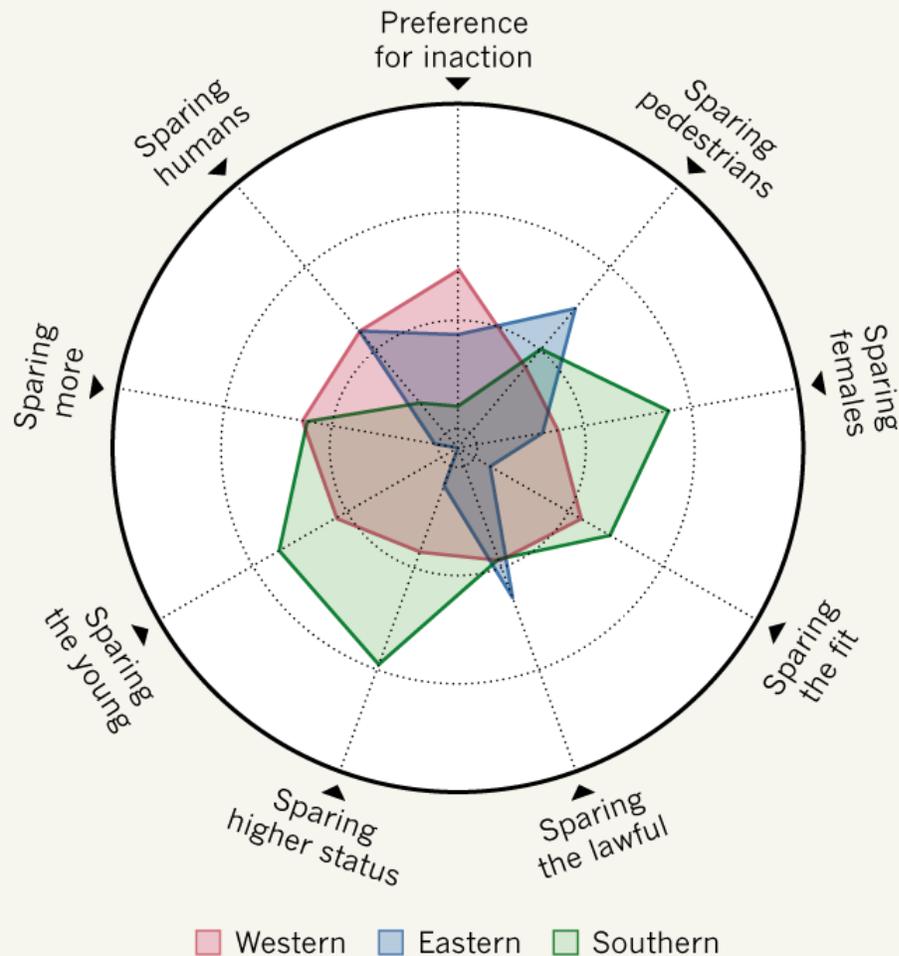
Bei unausweichlichen Unfallsituationen ist jede Qualifizierung nach persönlichen Merkmalen (Alter, Geschlecht, körperliche oder geistige Konstitution) strikt untersagt. Eine Aufrechnung von Opfern ist untersagt. Eine allgemeine Programmierung auf eine Minderung der Zahl von Personenschäden kann vertretbar sein. Die an der Erzeugung von Mobilitätsrisiken Beteiligten dürfen Unbeteiligte nicht opfern.

- The majority of members “liked” autonomous driving
- Kantian ethics dominated over utilitarianism



MORAL COMPASS

A survey of 2.3 million people worldwide reveals variations in the moral principles that guide drivers' decisions. Respondents were presented with 13 scenarios, in which a collision that killed some combination of passengers and pedestrians was unavoidable, and asked to decide who they would spare. Scientists used these data to group countries and territories into three groups based on their moral attitudes.



Les enfants d'Israël
commirent une
infidélité au sujet
des choses vouées
par interdit.

Josué déchire ses
vêtements, se
prosterne contre
terre et reste
devant l'arche de
l'Alliance jusqu'au
soir. *Js 7, 1-6*



Lorsque le Seigneur, béni soit-Il, dit à Josué: « Israël a péché », celui-ci Lui demanda : « Qui est celui qui a péché? »

**La réponse de l'Éternel : « Mais suis-je délateur ?
Va et jette les dés. »**

Talmud de Babylone, Sanhedrin 43b

Google is afraid of assuming your gender with Gmail's Smart Compose feature

TECHNOLOGY NEWS NOVEMBER 27, 2018

Autocomplete will not suggest gender-specific pronouns anymore

Google refused to take chances at a time when gender issues are reshaping politics and society.

New Message

Haven't heard about Pierre. How is it going?

Google decided the best solution was to **limit linguistic coverage.**

Alexei Grinbaum
LES ROBOTS ET LE MAL



DESCLÉE DE BROUWER