

# Ethische Herausforderungen von ChatGPT

Fribourg, 25. Mai 2023

Dr. Johan Rochel

# Disclaimer

ethix

Lab for Innovation Ethics  
Laboratoire d'éthique de l'innovation  
Laboratorio per l'etica dell'innovazione  
Lab für Innovationsethik

EPFL

UNI  
FR  
■

UNIVERSITÉ DE FRIBOURG  
UNIVERSITÄT FREIBURG

- + New chat
- AI in Education Challenges.
- New chat
- New chat
- New chat
- New chat
- New chat
- New chat
- New chat
- New chat
- Clear conversations
- Upgrade to Plus NEW
- Dark mode
- Updates & FAQ
- Log out

# ChatGPT



## Examples

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →



## Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests



## Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge of world and events after 2021

[ChatGPT Feb 13 Version](#). Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

# Identifizierte Anwendungen

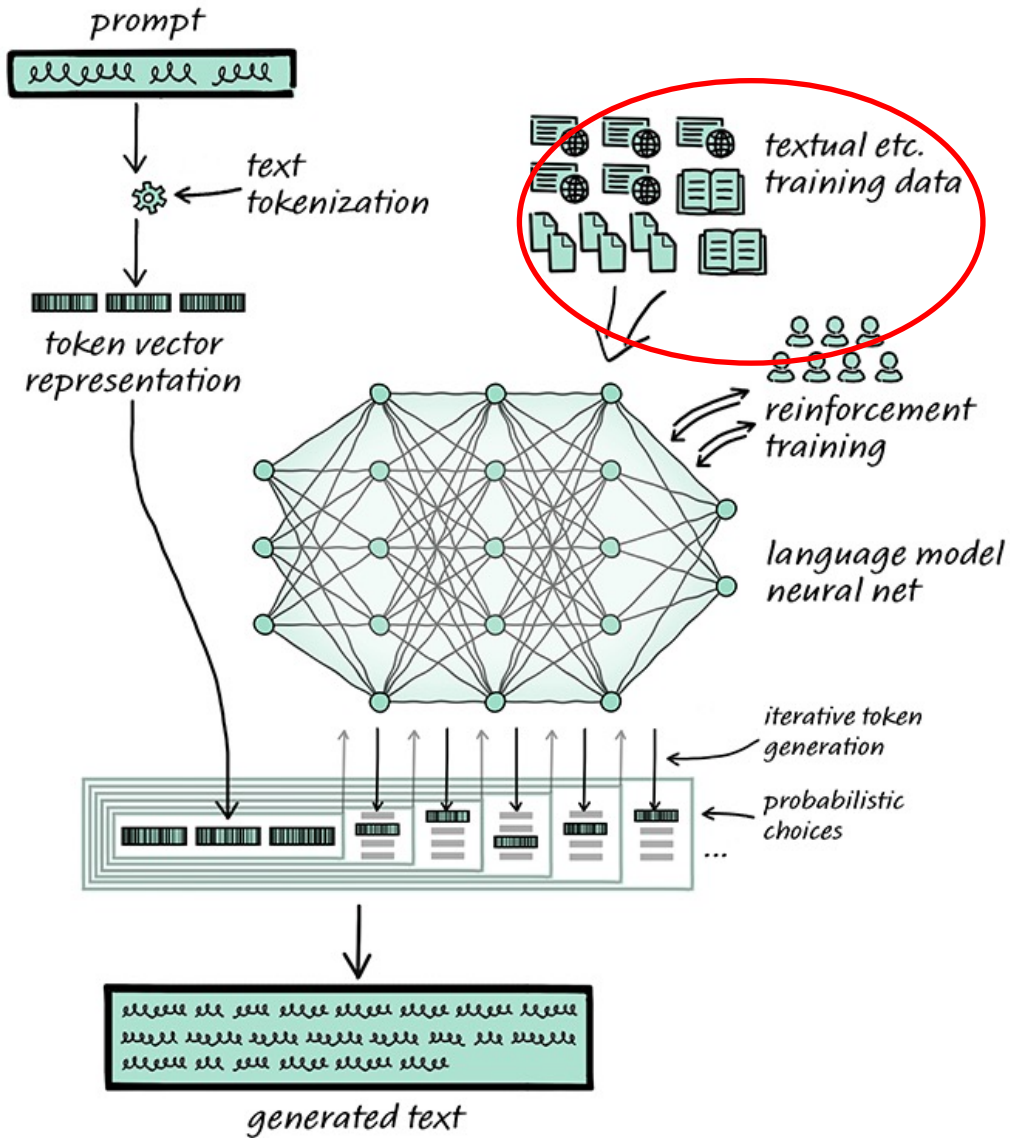
- **Erster Entwurf** eines Textes auf der Grundlage einiger Inputs
- **Stilistische Variation** eines bestimmten Textes
- **Sparringpartner**, um zu überprüfen, ob die wichtigsten Punkte zu einem bestimmten Problem identifiziert wurden
- **Zusammenfassung/Synthese** eines Textes auf der Grundlage von Bullet-points/Meetingnotizen
- **Grobschnitt** für eine erste Informationssuche

*(Coding-Tool)*

# Wie funktioniert ein Large Language Model (LLM) ?

Hi Sebastian,  
I hope that this email finds you well

# ChatGPT



## Trainingsdaten ChatGPT

Dataset	Quantity (tokens)	Weight in training mix
Common Crawl (filtered)	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

(Redirigé depuis [Gateau](#))

Cet article est une **ébauche** concernant la **gastronomie**.



Vous pouvez partager vos connaissances en l'améliorant (**comment ?**) selon les recommandations des [projets correspondants](#).

Consultez la liste des **tâches à accomplir** en [page de discussion](#).

Un **gâteau** est une **pâtisserie** préparée à partir d'une **pâte sucrée** cuite au four, généralement dans un moule<sup>1</sup>. Il peut être garni de **crème**, de **fruits**, de **chocolat** ou de **glacage**. Il se mange au **goûter**, ou à la fin du repas, au **dessert**. Le gâteau est généralement de forme ronde, carrée ou rectangulaire et plutôt plate.

En **Suisse romande**, le terme « **gâteau** » désigne usuellement toutes les formes de **tartes**, qu'elles soient sucrées ou salées (comme les quiches)<sup>2</sup>, et la définition ci-dessus y apparaît comme typiquement française.

## Exemples de types de **gâteaux** [ modifier | modifier le code ]

- **Gâteau marbré**



Un Chanteclair,

“ C'était sous le hangar de la charretterie que la table était dressée. Il y avait dessus quatre aloyaux, six fricassées de poulets, c veau à la casserole, trois gigots, et, au milieu, un joli cochon de lait rôti, flanqué de quatre andouilles à l'oseille. Aux angles, se dressait l'eau-de vie dans des carafes. Le cidre doux en bouteilles poussait s mousse épaisse autour des bouchons, et tous les verres, d'avance, avaient été remplis de vin jusqu'au bord. De grands plats de crème jaune, qui flottaient d'eux-mêmes au moindre choc de la table, présentaient, dessinés sur leur surface unie, les chiffres des nouveua époux en arabesques de nonpareille. On avait été chercher un pâtissier à Yvetot, pour les tourtes et les nougats. Comme il débuta dans le pays, il avait soigné les choses; et il apporta, lui-même, au dessert, une pièce montée qui fit pousser des cris. A la base, d'abord c'était un carré de carton bleu figurant un temple avec portiques, colonnades et statuettes de stuc tout autour dans des niches constellées d'étoiles en papier doré; puis se tenait au second étage donjon en **gâteau** de Savoie, entouré de menues fortifications en angélique, amandes, raisins secs, quartiers d'oranges; et enfin, sur plate-forme supérieure, qui était une prairie verte où il y avait des rochers avec des lacs de confitures et des bateaux en écailles de noisettes, on voyait un petit Amour, se balançant à une escarpolett de chocolat, dont les deux poteaux étaient terminés par deux bout de rose naturels, en guise de boules, au sommet.

Accueil > Recette de gâteaux maison

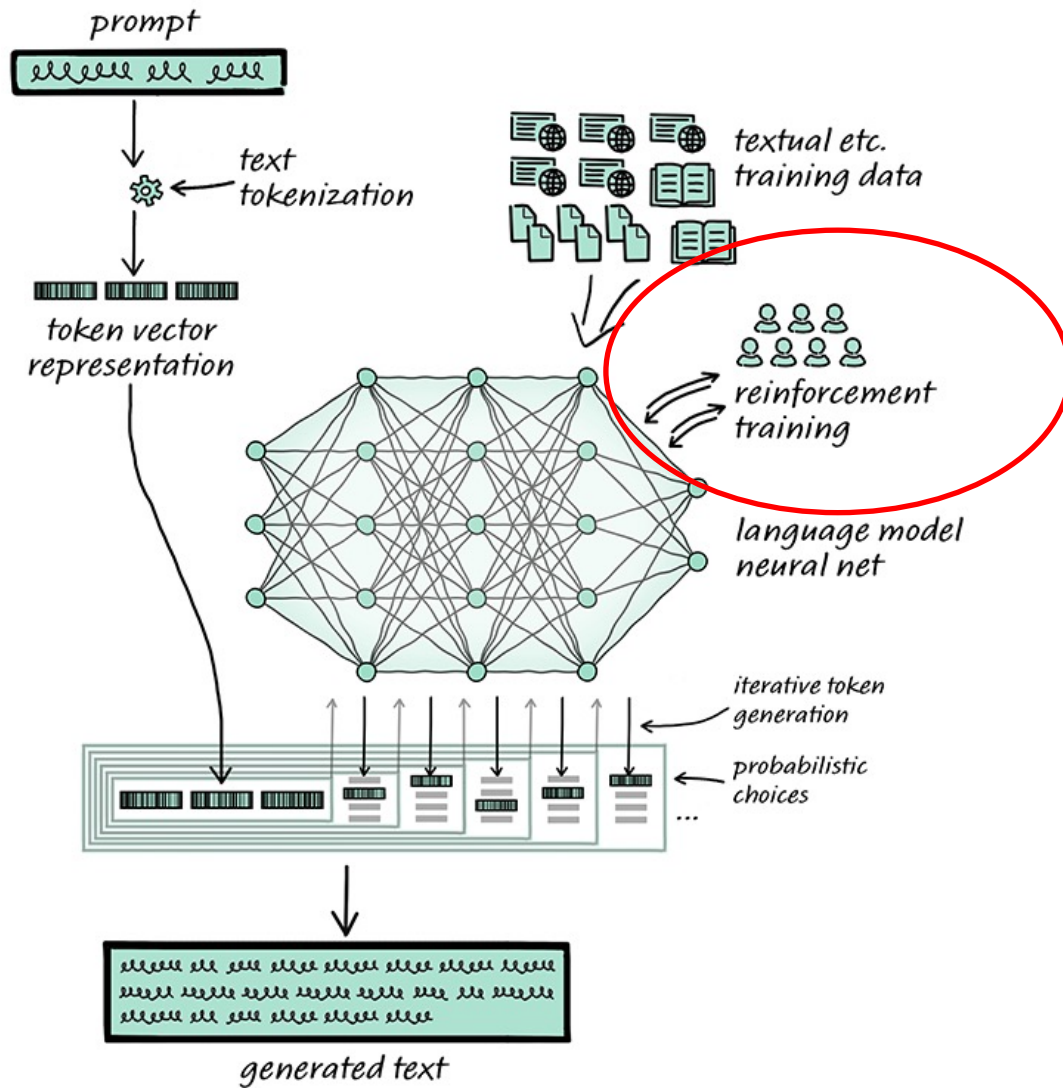
## Recettes de **gâteaux** maison

Découvrez nos meilleures recettes de **gâteaux** maison ! Pour un anniversaire, un goûter en famille, un pot de départ, un pique-nique, etc. Optez pour la simplicité en vous inspirant de nos nombreuses idées recettes de moelleux au chocolat faciles, **gâteaux** au yaourt express, quatre-quarts au beurre, et **gâteaux** aux pommes comme chez mamie. Vous pouvez aussi oser plus d'originalité en préparant un **gâteau** magique, un entremets comme chez le pâtissier, un somptueux gâteau à étages aux fruits frais et chantilly, ou encore un carrot cake, pour surprendre petits et grands gourmands. Il y en a pour tous les goûts. Vous avez enfilé votre tablier et préchauffé le four ? Parfait, à vous de jouer maintenant, en cuisine !





# ChatGPT



Step 1

Collect demonstration data and train a supervised policy.

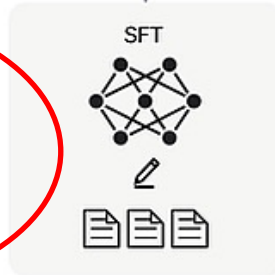
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



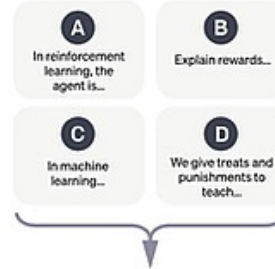
This data is used to fine-tune GPT-3.5 with supervised learning.



Step 2

Collect comparison data and train a reward model.

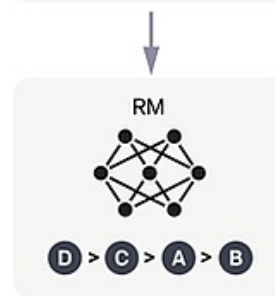
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

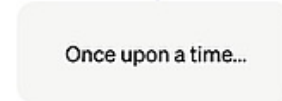
A new prompt is sampled from the dataset.



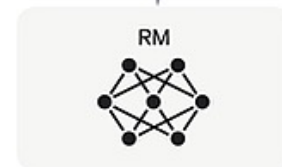
The PPO model is initialized from the supervised policy.



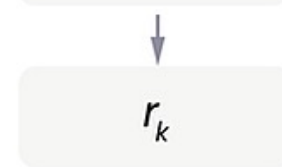
The policy generates an output.



The reward model calculates a reward for the output.



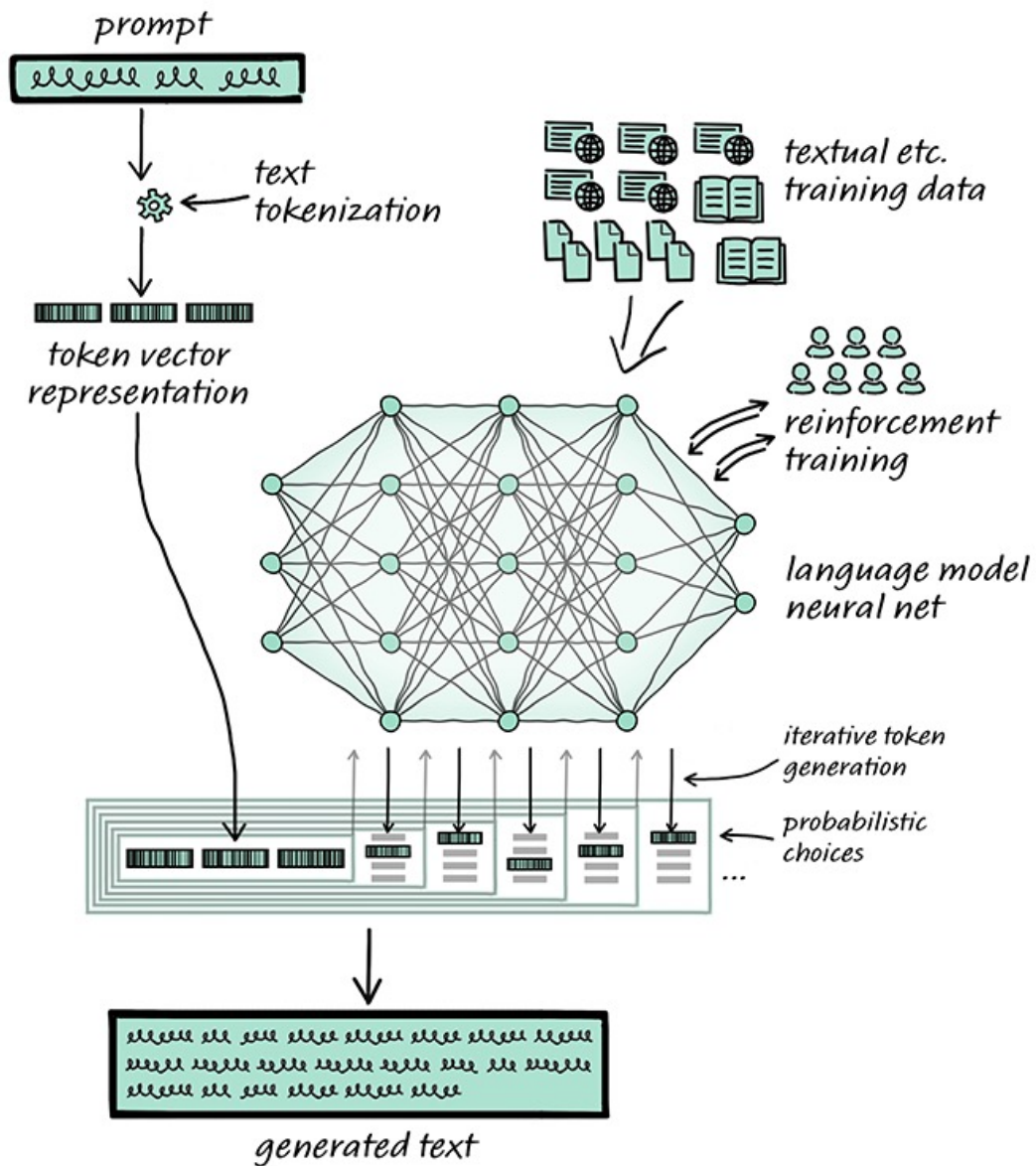
The reward is used to update the policy using PPO.



# Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic



# ChatGPT





wer ist christophe darbellay ?



Christophe Darbellay ist ein Schweizer Politiker und Mitglied der Schweizerischen Volkspartei (SVP). Er wurde 1971 in Siders im Kanton Wallis geboren. Von 2006 bis 2017 war er Mitglied des Nationalrates, dem schweizerischen Parlament. Darbellay war auch Mitglied des Walliser Gros



Où travaille Johan Rochel, l'éthicien bien connu ?



Je suis désolé, mais en tant que modèle de langage, je n'ai pas accès aux informations en temps réel sur les emplois ou les activités actuelles des personnes. Cependant, Johan Rochel est un éthicien connu pour ses travaux sur l'éthique de l'IA, de la robotique et des technologies émergentes. Il a travaillé dans différentes institutions académiques et organisations, notamment à l'Université de Twente aux Pays-Bas, à l'Université d'Oxford au Royaume-Uni, à la Fondation pour la recherche stratégique en France et à l'Organisation de coopération et de développement économiques (OCDE) à Paris.





# Öffentliche Debatte rund um ChatGPT

Ist ChatGPT zu menschlich?

Sagt ChatGPT die Wahrheit? Ist es neutral?

Ist ChatGPT kreativ?

Ist ChatGPT bewusstseinsfähig?  
Ist es intelligent?

## 6 questions éthiques-philosophiques

En libre accès, à utiliser pour  
enseignement/sensibilisation:

[https://ethix.ch/sites/default/files/inline-files/Ethix\\_ChatGPT\\_April2023.pdf](https://ethix.ch/sites/default/files/inline-files/Ethix_ChatGPT_April2023.pdf)





# Einige Empfehlungen

- Sich vor Anthropomorphisierung hüten - ChatGPT ist nur eine Ansammlung von mathematischen Formeln und Codezeilen.
- Einen kühlen Kopf bewahren - das Äussere vermittelt den Eindruck von Objektivität und Wahrhaftigkeit, aber der Inhalt kann falsch sein.
- Die Quellen des Inhalts angeben - wo sind die relevanten Informationen zu finden? Wie kann ich den «Black Box»-Effekt vermeiden?
- Fair Play spielen und die Ehrlichkeit meiner intellektuellen Arbeit sicherstellen - klarstellen und erklären, wie ich ChatGPT für meine Arbeit verwendet habe.

# Einige Empfehlungen

- Den Menschen hinter der KI suchen und die Frage nach seiner Verantwortung und seinen wirtschaftlichen Interessen stellen.
- Alle KIs als sozio-technische Systeme behandeln - sie sind immer in einen bestimmten Kontext eingebettet.
- Verhaltensweisen, die durch Technologie ermöglicht werden, als Fragen/Gegenstände politischer Entscheidungen behandeln.

# Drei Ebenen der digitalen Innovationsethik



# Drei Ebenen der digitalen Innovationsethik



## 1. Ethik der Technologie

Ethische Fragen zur **Technologie selbst**, u.a.:

- Datenethische Fragen: z.B. Verhältnismässigkeit, Transparenz, Einwilligung
- Algorithmenethische Fragen bzw. Ethik der KI: z.B. Nachvollziehbarkeit, Biases, Qualität von Trainingsdaten
- Designethik
- Nachhaltigkeit
- Kontrolle/Accountability

# Drei Ebenen der digitalen Innovationsethik



## 2. Soziale Gerechtigkeit

Auswirkungen eines Produkts/einer Dienstleistung/ eines Geschäftsmodells auf die **Gesellschaft**, u.a.:

- Art der Arbeit/Entwicklung Pflichtenheft
- Arbeit/Jobs und ihre Verteilung
- Sicherheit/Überwachung
- Chancengerechtigkeit
- Zugang zur Technologie

# Drei Ebenen der digitalen Innovationsethik



## 3. Auswirkungen auf Narrative

Auswirkungen einer Technologie/eines Produkts auf fundamentale **Framings und Narrative**, u.a.:

- Freiheit; (Willens-)Freiheit
- Mensch/Maschine und deren Interaktion
- Verhältnis Natur/Artefakt

# ChatGPT im Kontext der Schule

- Schritt 1: Was ermöglichen generative KIs heute?
- Schritt 2: Welchen Nutzen haben schriftliche Hausarbeiten in Bezug auf welche Kompetenzen?
- Schritt 3: Sollen diese Kompetenzen mit oder ohne technologische Unterstützung erreicht werden, und wenn ja, wer kann darüber entscheiden?
- Schritt 4: Welche unmittelbaren Risiken sind mit der Verwendung des Tools verbunden?
- Schritt 5: Was sind die längerfristigen Risiken?
- Schritt 6: Welche Anforderungen sind angesichts dieser Risiken an die Nutzung des Tools in einem Bildungskontext zu stellen?

# Anhang



# WARNING – Using GPT-tool to write paper

- Powerful tools have been made available (ChatGPT by OpenAI)
- It might be tempting to use them to write the term paper. Please consider the following points:
  - a. You might use these tools, as you might use other tools (such as “deepl” for translation purposes).
  - b. You bear the full responsibility for the text (its content, but also the correct referencing, according to the requirements of the course). You are being evaluated, not ChatGPT.
  - c. Please indicate whether you use such tools and, if yes, briefly discuss the following points in a specific annex (not part of the paper’s authorized length):
    - For which purposes have you been using these tools ? (eg. Provide a first draft ; sparring-partner to check whether I forgot something)
    - Which challenges do you see in using these tools ? Please write a few sentences of critical self-reflection.
    - What did you do to avoid these challenges ?

## Chinesisches Zimmer – J. Searle



# Für Interessierten

Weiterbildungskurs „Ethik, Digitalisierung und Innovation“

Ethik Zentrum UZH

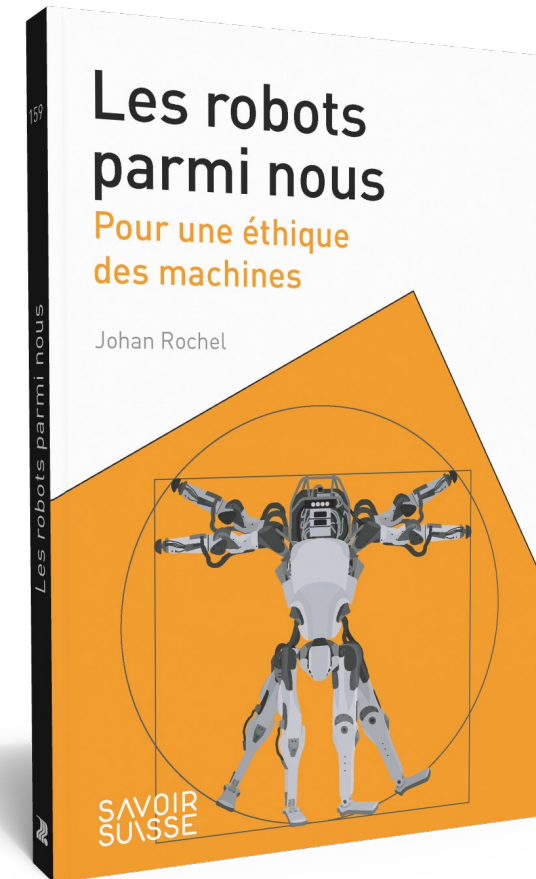
Freitag/Samstag, 24.-25. November 2023

<https://www.asae.uzh.ch/de/weiterbildungskurse/digitalisierungundinnovation.html>

*Les robots parmi nous – Pour une éthique des machines*

EPFL Press 2022

<https://www.epflpress.org/produit/1054/9782889154586/les-robots-parmi-nous>



# Quellen

Funktionieren ChatGPT: <https://writings.stephenwolfram.com/2023/01/wolframalpha-as-the-way-to-bring-computational-knowledge-superpowers-to-chatgpt/>

Quellen ChatGPT : <https://stanford-cs324.github.io/winter2022/lectures/data/#gpt-3-dataset>

Fine-tuning: <https://openai.com/research/instruction-following>

Gâteaux: <https://www.regal.fr/mots-cles/gateau> ; Flaubert « Madame Bovary »

Kenya: <https://time.com/6247678/openai-chatgpt-kenya-workers/>

Foto Junge Frauen: Miles, Twitter <https://twitter.com/mileszim>

Searle: Steven Novella, Stanford Encyclopedia: <https://plato.stanford.edu/entries/chinese-room/>

Vielen Dank!

[www.ethix.ch](http://www.ethix.ch)

[rochel@ethix.ch](mailto:rochel@ethix.ch)