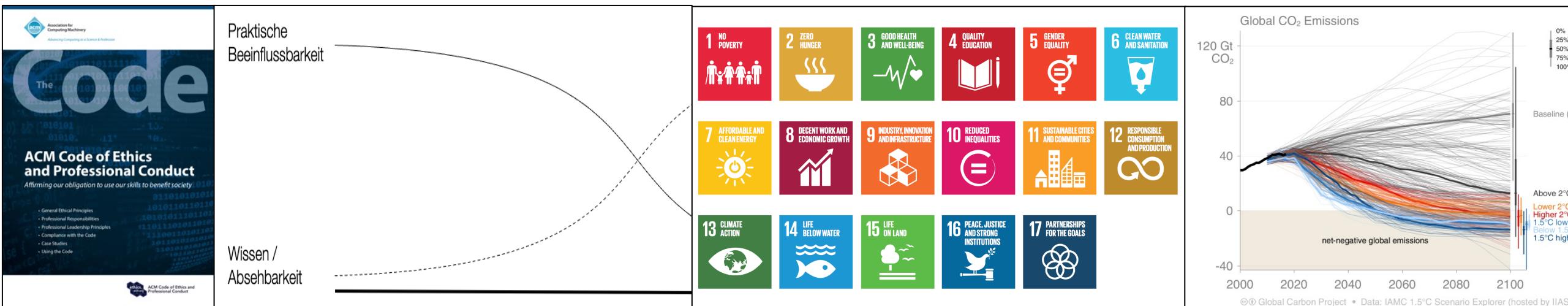


# Information Governance

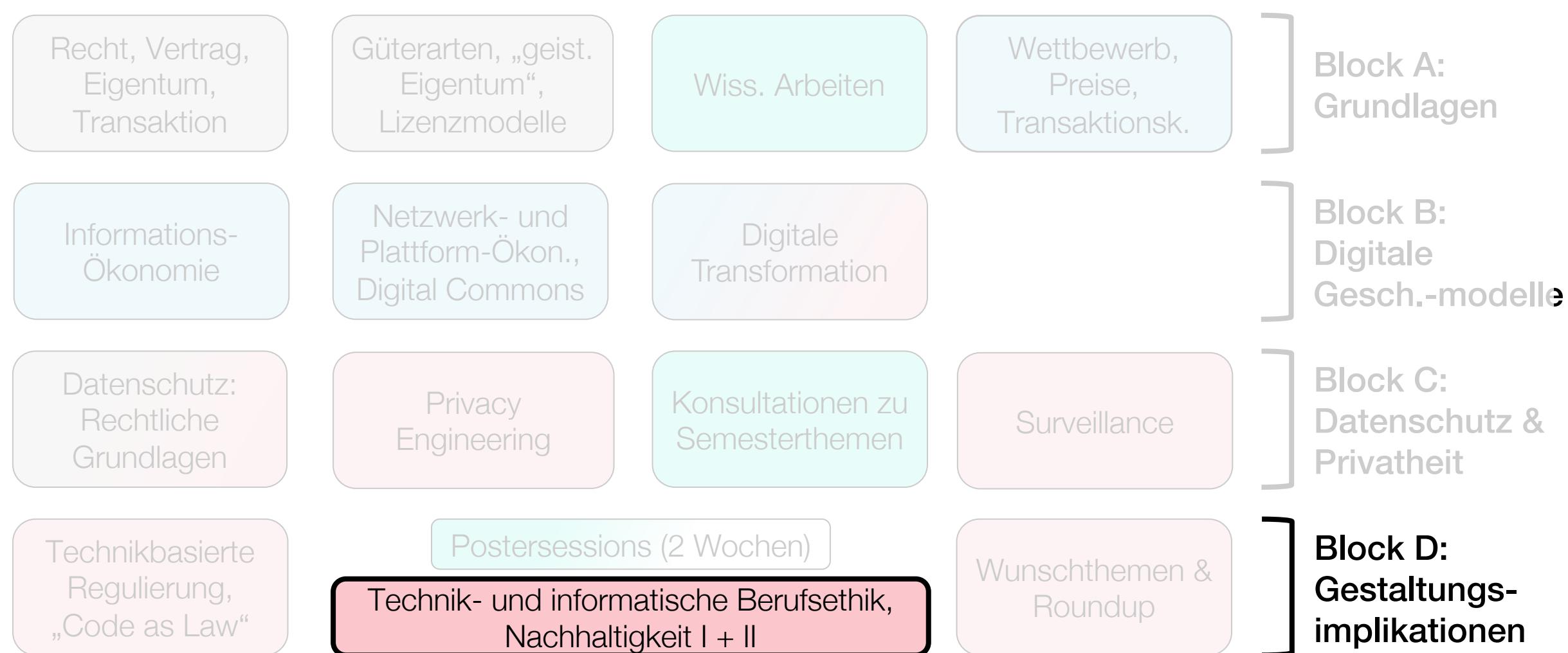
## Lesson 13: Technik-, Informations- und informatische Berufsethik & Nachhaltigkeit II



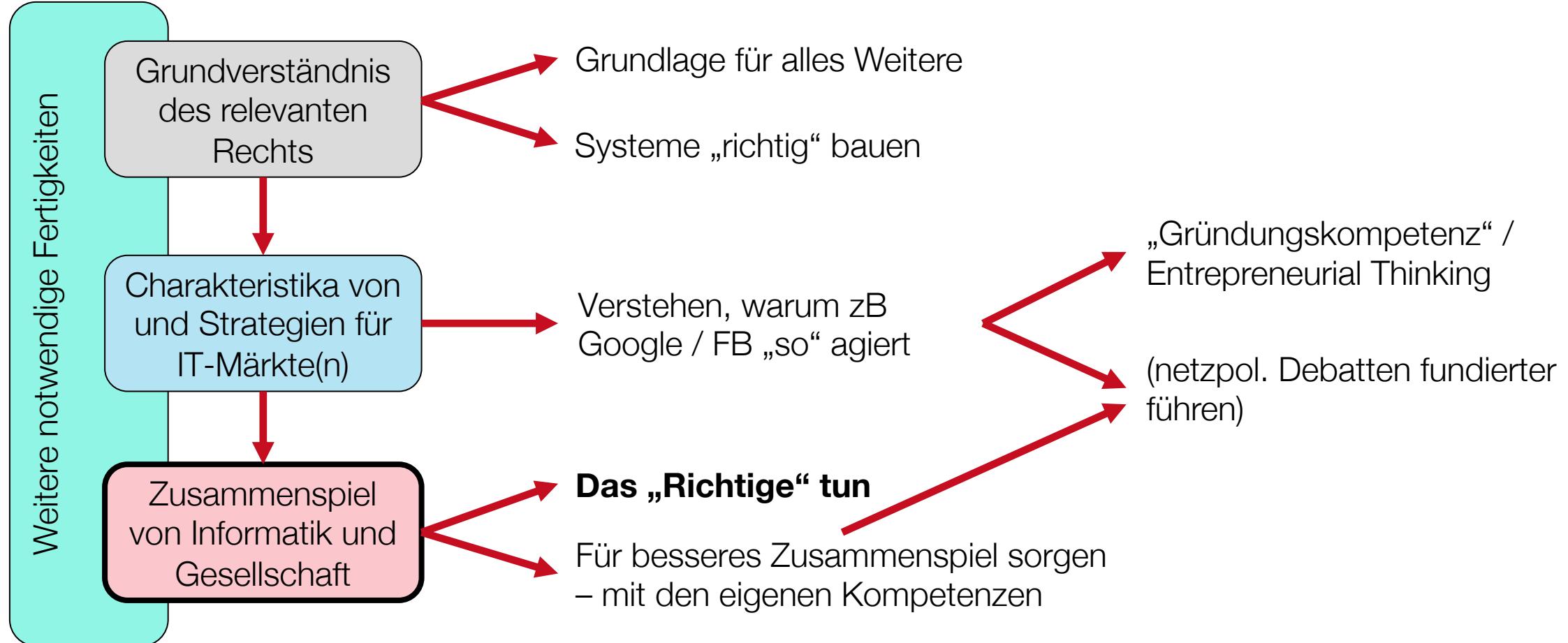
Frank Pallas/Elias Grünewald

*Information Systems Engineering*  
TU Berlin

# Information Governance – Thematischer Überblick



# Information Governance – „Riding Skills“



# Ethik I: Recap

**HAUPTBEITRAG / DAS MORALISCHE WISSEN VON STUDIERENDEN }**

*„Das muss man immer für sich selber abwägen“ oder:  
Das moralische Wissen von Studierenden der Informatik“*

Christoph Schneider

**„Technik bereitstellen oder nicht? Wenn das Informatik machen kann, ist ethisch, moralisch zuerst mal völlig neutral. Wie jemand dann die Technik anwendet, ist etwas anderes, hat aber mit uns nichts mehr zu tun.“**

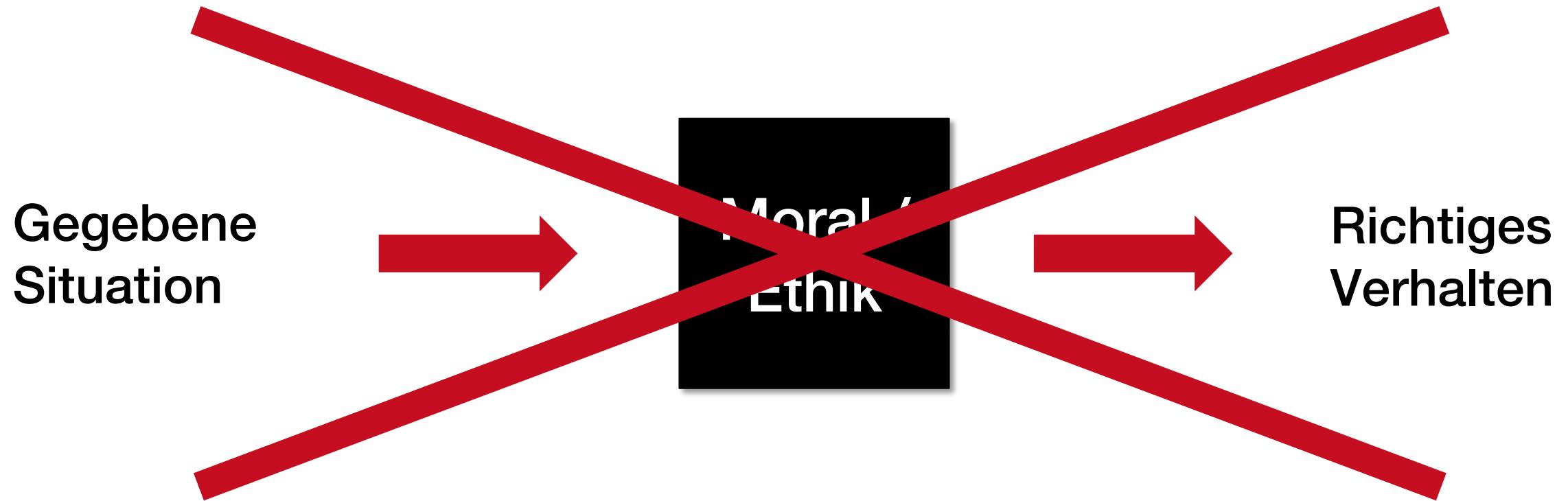
Ist Technikentwicklung wie so vieles ambivalent?  
Ohne endgültige Antworten zu liefern werde ich auf

Was halten Sie von nebenstehendem Statement? Stimmt? Stimmt nicht? Stimmt manchmal? Können Sie eindeutig Position beziehen? Muss man sich überhaupt ethische Fragen als InformatikerIn stellen?

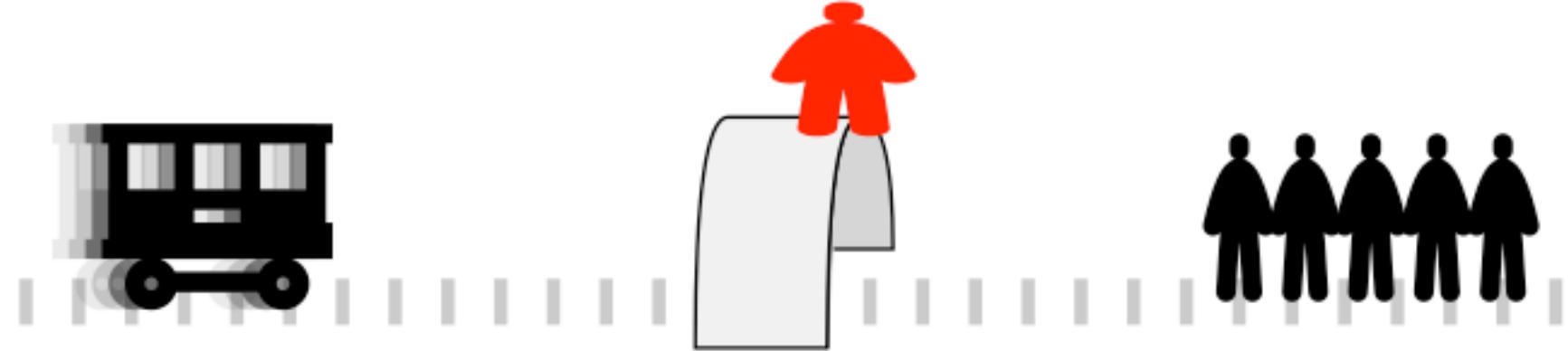
die Ohnmacht der Vernunft [14], hält Weizenbaum abschließend im zehnten Kapitel ein leidenschaftliches Plädoyer für eine universitäre Ausbildung von InformatikerInnen, die mehr ist als bloßes Training technischer Fähigkeiten. Statt nur zu erweitern was Computer können, müssen ihre EntwicklerInnen auch lernen zu bewerten, was Computer überhaupt tun sollen. Weizenbaum appelliert an die Dozierenden, durch ihr eigenes Beispiel den Studierenden die Gültigkeit eines weniger eindeutigen, ambivalenten moralischen Wissens, welches sich letzterem annimmt vorzuleben. Erst wenn man die ambivalente

<https://link.springer.com/content/pdf/10.1007%2Fs00287-013-0695-y.pdf>

# Ethik I: Recap



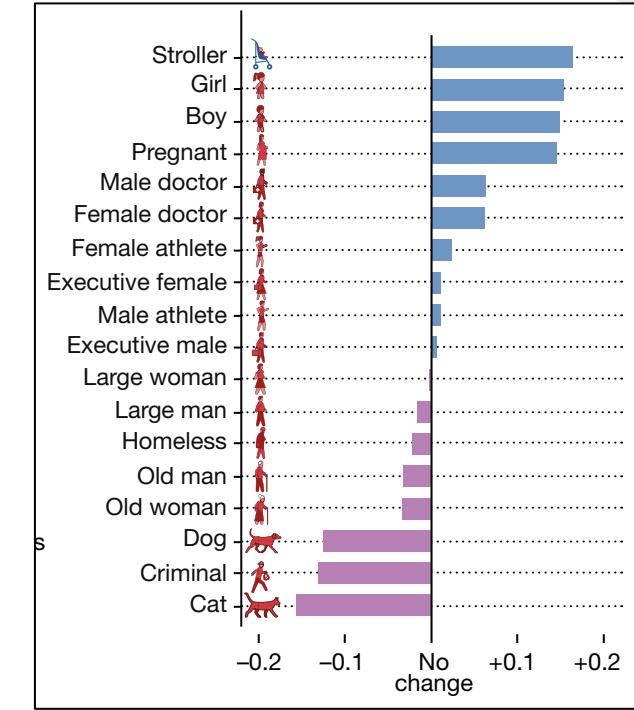
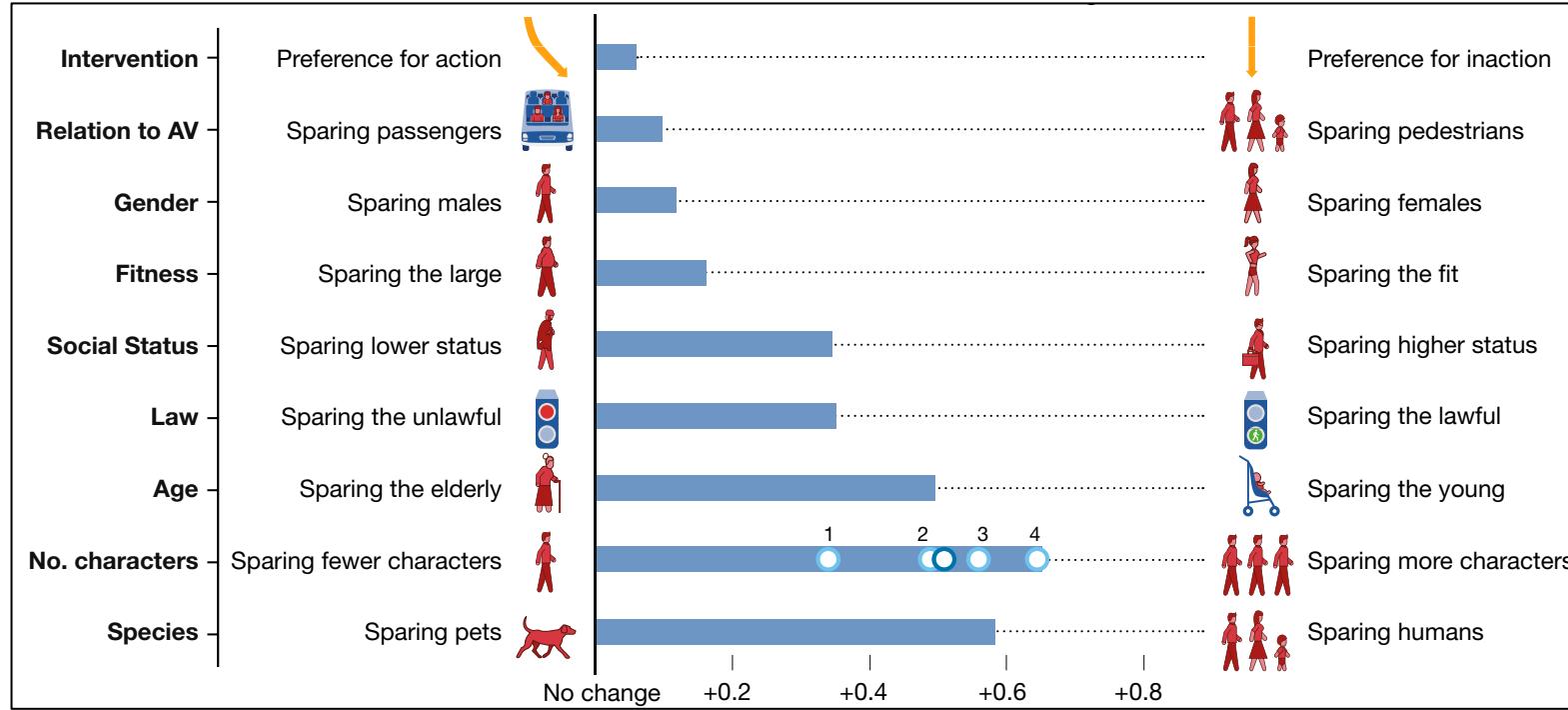
# Ethik I: Recap



Nichts tun → 5 Personen sterben

Sehr dicke\*n Verursacher\*in von der Brücke stoßen → 1 Person stirbt

# Ethik I: Recap



→ Befragte Menschen sind der Meinung, es sollte nach sozialem Status / Regelkonformität / Alter zu differenziert werden

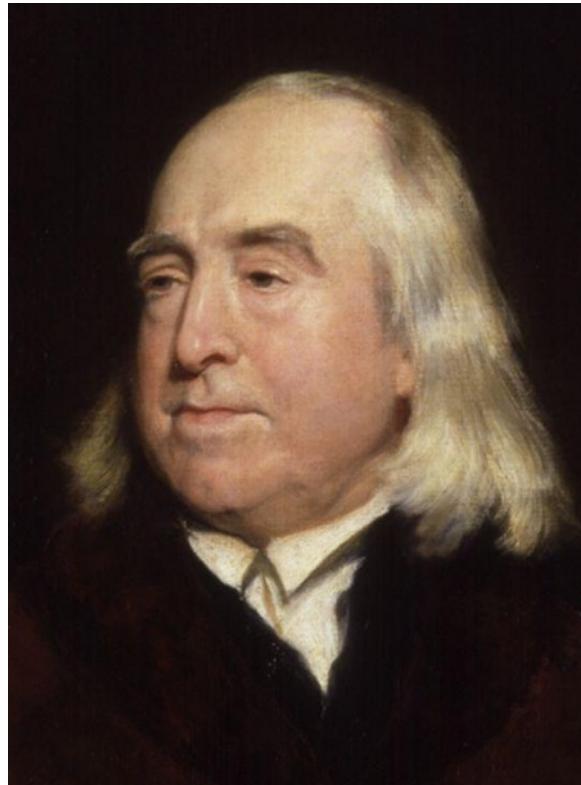
→ Befragte Menschen sind der Meinung, „Hunde zählen mehr als Kriminelle“

# Ethik I: Recap

„Ethik“ = „the study of morality“

Tavani (2013, 3)

# Ethik I: Recap



Jeremy Bentham – Utilitaristisch:  
„Greatest happiness of the  
greatest number“



Immanuel Kant – Deontologisch:  
„Der Mensch selbst als Ziel, nie als Mittel“  
Bestimmte Handlungen sind “per se” richtig  
bzw. falsch

usw. usf. → Unterschiedliche Ethiken können zu (signifikant) unterschiedlichen Bewertungen führen

# Ethics Shopping & Ethics Washing



I only realized that all this was not actually desired when our friendly Finnish HLEG President Pekka Ala-Pietilä (formerly Nokia) asked me in a gentle voice whether we could remove the phrase "non-negotiable" from the document. In the next step, many industry representatives and group members interested in a "positive vision" vehemently insisted that the phrase "Red Lines" be removed entirely from the text – although it was precisely these red lines that were our mandate. The published document no longer contains any talk of "Red Lines"; three were completely deleted and the rest were watered down. Instead there is only talk of "critical concerns".

<https://www.tagesspiegel.de/politik/e-u-guidelines-ethics-washing-made-in-europe/24195496.html>

This phenomenon is an example of "ethics washing". Industry organizes and cultivates ethical debates to buy time – to distract the public and to prevent or at least delay effective regulation and policy-making. Politicians also like to set up ethics committees because it gives them a course of action when, given the complexity of the issues, they simply don't know what to do – and that's only human. At the same time, however, industry is building one "ethics washing machine" after another. Facebook has invested in the TU Munich – funding an

Unterschiedliche „Ethiken“ mit unterschiedlichen Leitgedanken führen zu unterschiedlichen Ergebnissen

→ Risiko von „Feigenblatt-Ethik“ und „Ethics-Shopping“ zur Rechtfertigung im Vorhinein angestrebter Ergebnisse

→ Risiko von „Ethics-Washing“ zur Vermeidung starker, grundrechtsorientierter Regulierung

# Lesson 13: Technik-, Informations- und informatische Berufsethik & Nachhaltigkeit II

Eine spezifische „Informatik-Ethik“?

Möglichkeiten (und Grenzen) zur Einbindung in Entwicklungsprozesse

Ethisch motivierte Orientierungshilfen / Leitplanken

Nachhaltigkeit als verwandte Herausforderung

# Lesson 13: Technik-, Informations- und informatische Berufsethik & Nachhaltigkeit II

## Eine spezifische „Informatik-Ethik“?

Möglichkeiten (und Grenzen) zur Einbindung in Entwicklungsprozesse

Ethisch motivierte Orientierungshilfen / Leitplanken

Nachhaltigkeit als verwandte Herausforderung

# Besondere Gegebenheiten der Informatik

Kritische Systeme mit potenziell weitreichenden Auswirkungen

Schnelle Entwicklungszyklen, die kaum Raum für Folgenabschätzung lassen

Informatische Systeme “steuern” menschliches Verhalten  
→ Wohin soll Technik regeln?

“Autonome Systeme”

...

# Besondere Gegebenheiten der Informatik?



Wernher von Braun (wikimedia)



Robert Oppenheimer (wikimedia)

Spezifische Physik-Ethik?

# Besondere Gegebenheiten der Informatik

Kritische Systeme mit potenziell weitreichenden Auswirkungen

Schnelle Entwicklungszyklen, die kaum Raum für Folgenabschätzung lassen

Informatische Systeme “steuern” menschliches Verhalten

→ Wohin soll Technik regeln?

“Autonome Systeme”

...

→ **Keine eigene informatische Ethik, aber „typische“  
Informatische Sachverhalte / Dilemmata**

# „Typische“ informatische Sachverhalte / Dilemmata



- KI-Einsatz in der Medizin
  - Forschungsunterstützung
  - Safeguards in klinischen Studien
  - Generell: Risiko von „Automation Bias“ – was automatisiert „entschieden“ wird, wird wenig hinterfragt
- KI-Einsatz in der Bildung
  - ...
- ...

Empfehlungen wie etwa:

- Zertifizierungen und Zulassungsbehörden
- Auswahl von Datensätzen
- Designs gegen Automation Bias (→ „nudge“)
- Nutzung kann auch klar geboten sein
- ...

# Besondere Gegebenheiten der Informatik

Kritische Systeme mit potenziell weitreichenden Auswirkungen

Schnelle Entwicklungszyklen, die kaum Raum für Folgenabschätzung lassen

Informatische Systeme “steuern” menschliches Verhalten

→ Wohin soll Technik regeln?

“Autonome Systeme”

...

→ Keine eigene informatische Ethik, aber „typische“ Informatische Sachverhalte / Dilemmata → Fallspezifische Anwendung etablierter Theorien!

# Lesson 13: Technik-, Informations- und informatische Berufsethik & Nachhaltigkeit II

Eine spezifische „Informatik-Ethik“?

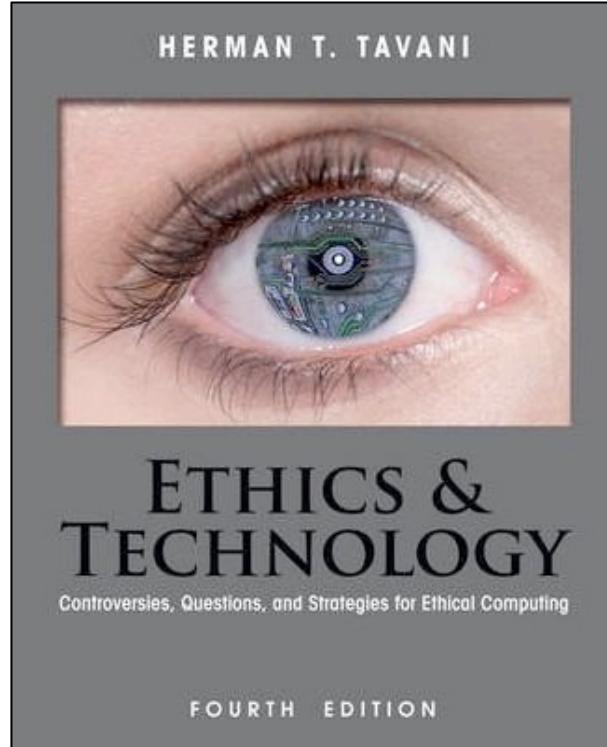
**Möglichkeiten (und Grenzen) zur Einbindung in Entwicklungsprozesse**

Ethisch motivierte Orientierungshilfen / Leitplanken

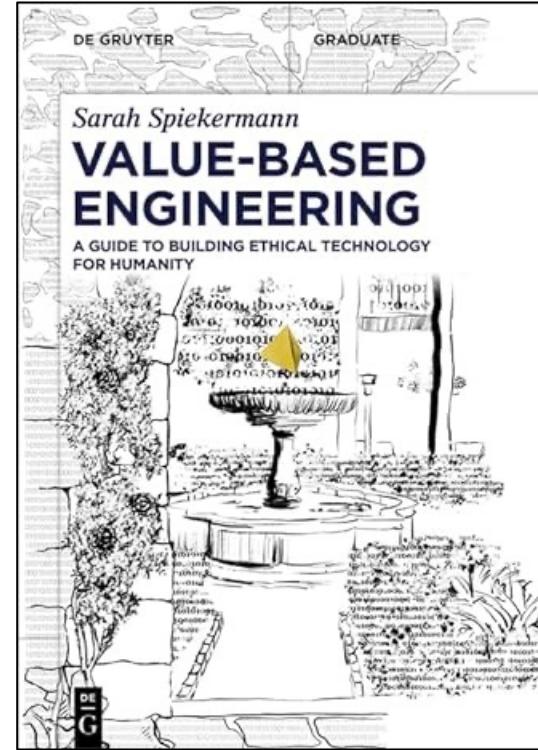
Nachhaltigkeit als verwandte Herausforderung

Was tun?

# Einbindung von Ethik in Technikgestaltung



Tavani



Spiekermann



## 3-Step-Process (Tavani 2013, 111f)

- Step 1: Identify a practice involving cybertechnology, or a feature of that technology, that is controversial from a moral perspective
  - 1a: **Disclose any hidden or opaque features**
  - 1b: **Assess** any descriptive components of the ethical issue via the **sociological implications** it has for relevant social institutions and sociodemographic groups
  - 1c: In analyzing normative elements of that issue, determine whether there are any guidelines [...] that can help resolve that issue [...]
  - If the normative ethical issue cannot be resolved through the application of existing policies [...] go to step 2

## 3-Step-Process (Tavani 2013, 111f)

- Step 1: Identify a practice involving cybertechnology, or a feature of that technology, that is controversial from a moral perspective
- Step 2: Analyze the ethical issue by **clarifying concepts and situating it in a context**
  - 2a: If a policy vacuum exists, go to step 2b; otherwise, go to step 3
  - 2b: Clear up any conceptual muddles involving the policy vacuum and go to Step 3

## 3-Step-Process (Tavani 2013, 111f)

- Step 1: Identify a practice involving cybertechnology, or a feature of that technology, that is controversial from a moral perspective
- Step 2: Analyze the ethical issue by clarifying concepts and situating it in a context
- Step 3: Deliberate on the ethical issue. The deliberation process requires two stages:
  - 3a: **Apply one or more ethical theories** [...] to the analysis of the moral issue [...]
  - 3b: Justify the position you reached [through] logic argumentation

Problem(e)?

## 3-Step-Process (Tavani 2013, 111f) – Challenge

- Step 1: Identify a practice involving cybertechnology, or a feature of that technology, that is controversial from a moral perspective
  - 1a: **Disclose** any hidden or opaque **features**
  - 1b: **Assess** any descriptive components of the ethical issue via the **sociological implications** it has for relevant social institutions and sociodemographic groups
  - 1c: In analyzing normative elements of that issue, determine whether there are any guidelines [...] that can help resolve that issue [...]
  - If the normative ethical issue cannot be resolved through the application of existing policies [...] go to step 2

# Besondere Gegebenheiten der Informatik

Kritische Systeme mit potenziell weitreichenden Auswirkungen

**Schnelle Entwicklungszyklen, die kaum Raum für Folgenabschätzung lassen**

Informatische Systeme “steuern” menschliches Verhalten

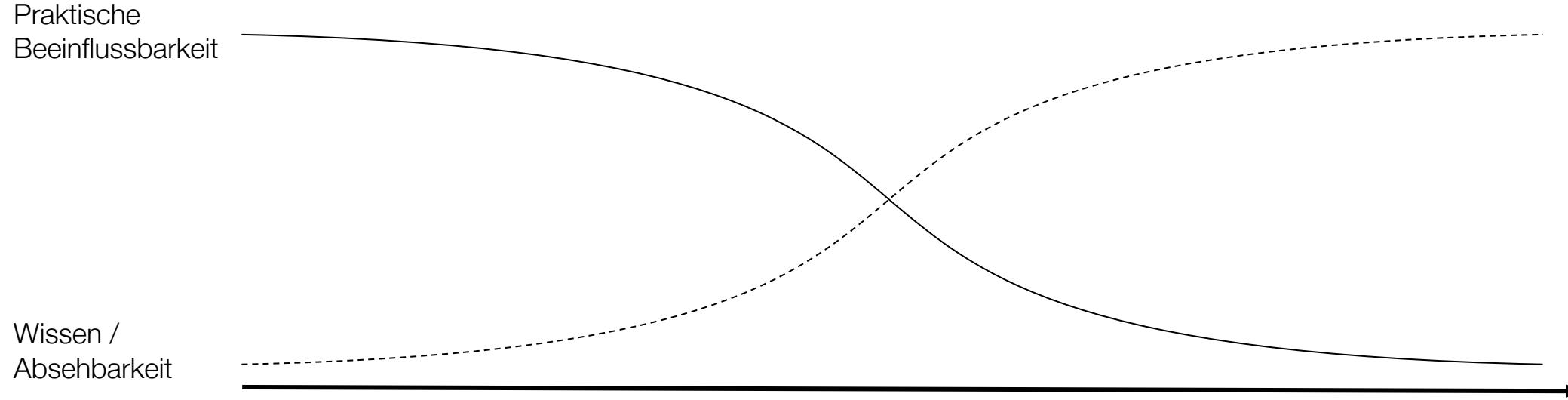
→ Wohin soll Technik regeln?

“Autonome Systeme”

...

→ **Keine eigene informatische Ethik, aber „typische“  
Informatische Sachverhalte / Dilemmata**

# Folgenabschätzung: Collingridge-Dilemma



In frühen Entwicklungsstadien ist Einfluss gut möglich, aber Folgen/Entwicklungsfade schwer abschätz-/vorhersehbar

In späteren Stadien werden Folgen klarer, aber Einfluss ist nur noch schwerlich möglich  
(direkte Kosten der Umsteuerung, sozio-technische lock-ins, ...)

## 3-Step-Process (Tavani 2013, 111f)

- Step 1: Identify a practice involving cybertechnology, or a feature of that technology, that is controversial from a moral perspective
  - 1a: Disclose any hidden or opaque features
  - 1b: Assess any descriptive components of the ethical issue via the sociological implications it has for relevant social institutions and sociodemographic groups
  - 1c: In analyzing normative elements of that issue, **determine whether there are any guidelines** [...] that can help resolve that issue [...]
  - If the normative ethical issue cannot be resolved through the application of existing policies [...] go to step 2

# Lesson 13: Technik-, Informations- und informatische Berufsethik & Nachhaltigkeit II

Eine spezifische „Informatik-Ethik“?

Möglichkeiten (und Grenzen) zur Einbindung in Entwicklungsprozesse

**Ethisch motivierte Orientierungshilfen / Leitplanken**

Nachhaltigkeit als verwandte Herausforderung

# „Informatische Berufsethik“: IEEE Code of Ethics



# „Informatische Berufsethik“: IEEE Code of Ethics

We, the members of the IEEE [...] agree:

1. to hold paramount the safety, health, and welfare of the public, to strive to comply with ethical design and sustainable development practices, to protect the privacy of others, and to disclose promptly factors that might endanger the public or the environment;
2. to improve the understanding by individuals and society of the capabilities and societal implications of conventional and emerging technologies, including intelligent systems;
3. to avoid real or perceived conflicts of interest whenever possible, and to disclose them to affected parties when they do exist;
4. to avoid unlawful conduct in professional activities, and to reject bribery in all its forms;
5. to seek, accept, and offer honest criticism of technical work, to acknowledge and correct errors, to be honest and realistic in stating claims or estimates based on available data, and to credit properly the contributions of others;
6. to maintain and improve our technical competence and to undertake technological tasks for others only if qualified by training or experience, or after full disclosure of pertinent limitations;
7. to treat all persons fairly and with respect, and to not engage in discrimination based on characteristics such as race, religion, gender, disability, age, national origin, sexual orientation, gender identity, or gender expression;
8. to not engage in harassment of any kind, including sexual harassment or bullying behavior;
9. to avoid injuring others, their property, reputation, or employment by false or malicious actions, rumors or any other verbal or physical abuses;
10. to support colleagues and co-workers in following this code of ethics, to strive to ensure the code is upheld, and to not retaliate against individuals reporting a violation.

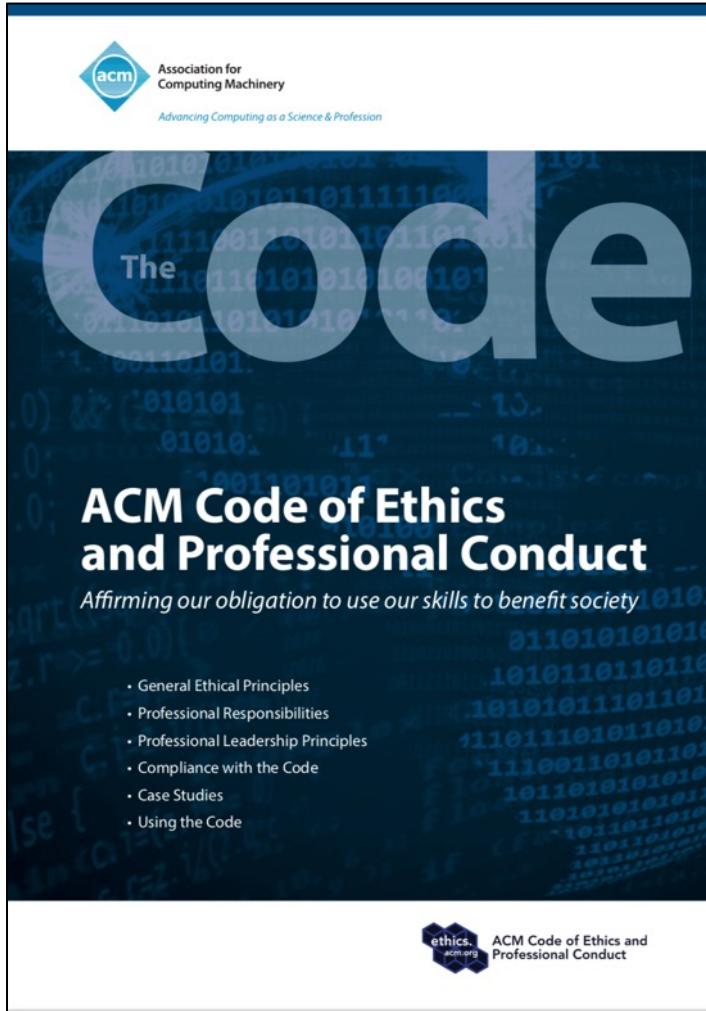
# „Informatische Berufsethik“: IEEE Code of Ethics

We, the members of the IEEE [...] agree:

1. to hold paramount the safety, health, and welfare of the public, [...] to protect the privacy of others, [...]
2. to improve the understanding [...] of conventional and emerging technologies, [...]
3. to avoid real or perceived conflicts of interest [...]
4. to avoid unlawful conduct in professional activities, and to reject bribery [...]
5. to seek, accept, and offer honest criticism of technical work [...]
6. to maintain and improve our technical competence [...]
7. to treat all persons fairly and with respect [...]
8. to not engage in harassment of any kind, [...]
9. to avoid injuring others, their property, reputation, or employment by false or malicious actions [...]
10. to support colleagues and co-workers in following this code of ethics, [...]

→ Bis auf „privacy“ kaum spezifisch für Informatik-Kontext  
„passt auch für Maschinenbau“)

# „Informatische Berufsethik“: ACM



## General Ethical Principles:

- Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing
- Avoid harm
- Be honest and trustworthy
- Be fair and take action not to discriminate
- Respect the work required to produce new ideas, inventions, creative works, and computing artifacts
- Respect privacy
- Honor confidentiality

<https://www.acm.org/code-of-ethics>

# „Informatische Berufsethik“: ACM

## Professional Responsibilities

- Strive to achieve high quality in both the processes and products of professional work
- Maintain high standards of professional competence, conduct, and ethical practice
- Know and respect existing rules pertaining to professional work
- Accept and provide appropriate professional review
- Give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks
- Perform work only in areas of competence
- Foster public awareness and understanding of computing, related technologies, and their consequences
- Access computing and communication resources only when authorized or when compelled by the public good
- Design and implement systems that are robustly and usably secure

# „Informatische Berufsethik“: ACM

## Professional Leadership Principles

- Ensure that the public good is the central concern during all professional computing work
- Articulate, encourage acceptance of, and evaluate fulfillment of social responsibilities by members of the organization or group
- Manage personnel and resources to enhance the quality of working life
- Articulate, apply, and support policies and processes that reflect the principles of the Code
- Create opportunities for members of the organization or group to grow as professionals
- Use care when modifying or retiring systems
- Recognize and take special care of systems that become integrated into the infrastructure of society

→ Insgesamt auch sehr nah am klassischen „Ingenieur:innen-Ethos“ (wie IEEE)

# „Informatische Berufsethik“: GI



Die Gesellschaft für Informatik e.V. (GI) will mit diesen Leitlinien bewirken, dass berufsethische oder moralische Konflikte Gegenstand gemeinsamen Nachdenkens und Handelns werden. Die Leitlinien sollen den GI-Mitgliedern und darüber hinaus allen Menschen, die IT-Systeme entwerfen, herstellen, betreiben oder verwenden, eine Orientierung bieten.

[...]

**Die GI und ihre Mitglieder verpflichten sich** zur Einhaltung dieser Leitlinien. Sie wirken auch außerhalb der GI darauf hin, dass diese im öffentlichen Diskurs Beachtung finden.

[...]

# „Informatische Berufsethik“: GI



- Fachkompetenz
- Sachkompetenz und kommunikative Kompetenz
- Juristische Kompetenz
- Urteilsfähigkeit
- Arbeitsbedingungen
- Organisationsstrukturen
- Lehren und Lernen
- Forschung
- Zivilcourage
- Soziale Verantwortung
- Ermöglichung der Selbstbestimmung

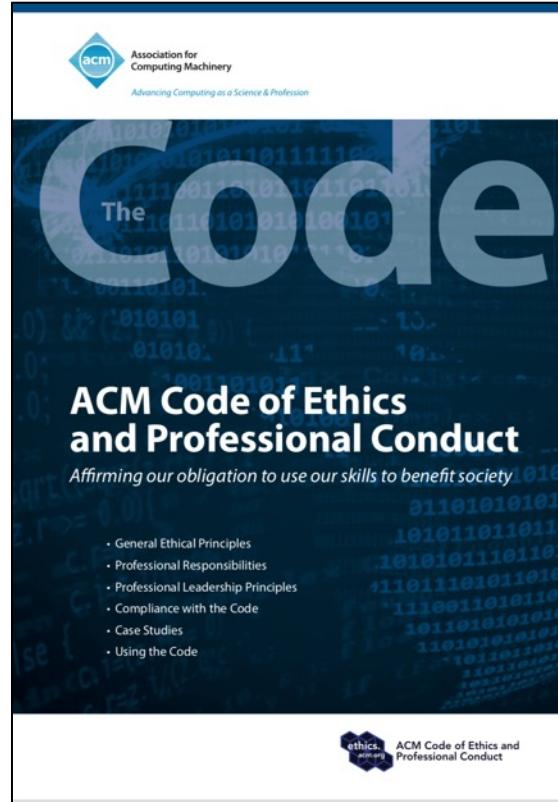
→ *Die Gesellschaft für Informatik ermutigt ihre Mitglieder, sich in jeder Situation an den Leitlinien zu orientieren. In Konfliktfällen versucht die GI zwischen den Beteiligten zu vermitteln.*

# „Informatische Berufsethik“?

- Meist eher Moral / Berufsehre
- „Codes of Ethics“ wären daher besser als „Codes of Conduct“ o.ä. bezeichnet

Was tun?

# „Leitfäden“ zu verantwortungsvollem Berufshandeln



→ Verschaffen Gefühl für informatische „Berufsmoral“

# „Leitfäden“ zur verantwortungsvollen Systemgestaltung

## Ethics for the New Surveillance

Gary T. Marx

Department of Sociology, University of Colorado, Boulder, Colorado, USA

The Principles of Fair Information Practice are almost three decades old and need to be broadened to take account of new technologies for collecting personal information such as drug testing, video cameras, electronic location monitoring, and the Internet. I argue that the ethics of a surveillance activity must be judged according to the means, the context and conditions of data collection, and the uses/goals, and suggest 29 questions related to this. The more one can answer these questions in a way that affirms the underlying principle (or a condition supportive of it), the more ethical the use of a tactic is likely to be. Four conditions are identified that, when breached, are likely to violate an individual's reasonable expectation of privacy. Respect for the dignity of the person is a central factor and emphasis is put on the avoidance of harm, validity, trust, notice, and permission when crossing personal borders.

**Keywords** borders, ethics, new information technologies, privacy, reasonable expectation of privacy, surveillance

"If it doesn't look right, that's ethics."  
—Popular expression

Received 1 April 1997; accepted 17 December 1997.

This article extends a paper delivered at the 1996 University of Victoria conference on Visions of Privacy in the Twenty-First Century. It is part of a broader project based on the Jensen Lectures delivered at Duke University, which will eventually appear in *Windows Into the Soul: Surveillance and Society in an Age of High Technology*. I am grateful to Hugo Bedau, Richard Leo, Helen Nissenbaum, Greg Ungar, Mary Virochow, and Lois Weithorn for their critical reading and suggestions. The paper was prepared while the author was a Fellow at the Center for Advanced Study in the Behavioral Sciences. I am grateful for financial support provided by the National Science Foundation grant SBR-9022192.

Address correspondence to Gary T. Marx, E-mail: Gary.Marx@Colorado.edu. Web: <http://socsci.colorado.edu/~marx/garyhome.html>

The Information Society, 14:171–185, 1998  
Copyright © 1998 Taylor & Francis  
0197-2243/98 \$12.00 + .00

171

"I'm in computer science. I took this class because eventually I want to do the right thing."  
—M.I.T. student

"It's a remarkable piece of apparatus."  
—F. Kafka, *The Penal Colony*

In 1928 Justice Brandeis wrote, "Discovery and invention have made it possible for the government, by means far more effective than stretching upon the rack, to obtain disclosure in court of what is whispered in the closet. The progress of science in furnishing the government with means of espionage is not likely to stop with wiretapping" (*Olmstead v. United States*, 1928). His haunting and prescient words clearly apply today, as the line between science and science fiction is continually redrawn and private-sector data collection practices join those of government as a cause of concern.

New technologies for collecting personal information that transcend the physical, liberty-enhancing limitations of the old means are constantly appearing. They probe more deeply, widely, and softly than traditional methods, transcending barriers (whether walls, distance, darkness, skin, or time) that historically made personal information inaccessible. The boundaries that have defined and given integrity to social systems, groups, and the self are increasingly permeable. The power of governmental and private organizations to compel disclosure (whether based on law or circumstance) and to aggregate, analyze, and distribute personal information is growing rapidly.

We are becoming a transparent society of record such that documentation of our history, current identity, location, physiological and psychological states, and behavior is increasingly possible. With predictive profiles there are even claims to be able to know individual futures. Information collection often occurs invisibly, automatically, and remotely, being built into routine activities. Awareness and genuine consent on the part of the subject may be lacking. The amount of personal information collected is

**TABLE 1**  
Questions to Help Determine the Ethics of Surveillance

- |  |
|--|
| <p>A. The Means</p> <ol style="list-style-type: none"> <li>1. Harm: Does the technique cause unwarranted physical or psychological harm?</li> <li>2. Boundary: Does the technique cross a personal boundary without permission (whether involving coercion or deception or a body, relational, or spatial border)?</li> <li>3. Trust: Does the technique violate assumptions that are made about how personal information will be treated, such as no secret recordings?</li> <li>4. Personal relationships: Is the tactic applied in a personal or impersonal setting?</li> <li>5. Invalidity: Does the technique produce invalid results?</li> </ol> <p>B. The Data Collection Context</p> <ol style="list-style-type: none"> <li>6. Awareness: Are individuals aware that personal information is being collected, who seeks it, and why?</li> <li>7. Consent: Do individuals consent to the data collection?</li> <li>8. Golden rule: Would those responsible for the surveillance (both the decision to apply it and its actual application) agree to its subjects under the conditions in which they apply it to others?</li> <li>9. Minimization: Does a principle of minimization apply?</li> <li>10. Public decision-making: Was the decision to use a tactic arrived at through some public discussion and decision-making process?</li> <li>11. Human review: Is there human review of machine-generated results?</li> <li>12. Right of inspection: Are people aware of the findings and how they were created?</li> <li>13. Right to challenge and express a grievance: Are there procedures for challenging the results, or for entering alternative data or interpretations into the record?</li> <li>14. Redress and sanctions: If the individual has been treated unfairly and procedures violated, are there appropriate means of redress? Are there means for discovering violations and penalties to encourage responsible surveillance behavior?</li> <li>15. Adequate data stewardship and protection: Can the security of the data be adequately protected?</li> <li>16. Equality-inequality regarding availability and application:           <ol style="list-style-type: none"> <li>(a) Is the means widely available restricted to only the most wealthy, powerful, or technologically sophisticated?</li> <li>(b) Within a setting is the tactic broadly applied to all people or only to those less powerful or unable to resist?</li> <li>(c) If there are means of resisting the provision of personal information are these means equally available, or restricted to the most privileged?</li> </ol> </li> <li>17. The symbolic meaning of a method: What does the use of a method communicate more generally?</li> <li>18. The creation of unwanted precedents: Is it likely to create precedents that will lead to its application in undesirable ways?</li> <li>19. Negative effects on surveillancees and third parties: Are there negative effects on those beyond the subject and, if so, can they be adequately mediated?</li> </ol> <p>C. Uses</p> <ol style="list-style-type: none"> <li>20. Beneficiary: Does application of the tactic serve broad community goals, the goals of the object of surveillance, or the personal goals of the data collector?</li> <li>21. Proportionality: Is there an appropriate balance between the importance of the goal and the cost of the means?</li> <li>22. Alternative means: Are other, less costly means available?</li> <li>23. Consequences of inaction: Where the means are very costly, what are the consequences of taking no surveillance action?</li> <li>24. Protections: Are adequate steps taken to minimize costs and risk?</li> <li>25. Appropriate vs. inappropriate goals: Are the goals of the data collection legitimate?</li> <li>26. The goodness of fit between the means and the goal: Is there a clear link between the information collected and the goal sought?</li> <li>27. Information used for original vs. other unrelated purposes: Is the personal information used for the reasons offered for its collection and for which consent may have been given, and do the data stay with the original collector, or do they migrate elsewhere?</li> <li>28. Failure to share secondary gains from the information: Is the personal data collected used for profit without permission from, or benefit to, the person who provided it?</li> <li>29. Unfair disadvantage: Is the information used in such a way as to cause unwarranted harm or disadvantage to its subject?</li> </ol> |
|--|

→ Helfen dabei, jedenfalls keine fundamentalen Aspekte zu übersehen  
(Bergen aber auch die üblichen Probleme von "Checklisten")

# „Beispielfälle“ spezifisch informatischer Dilemmata

The screenshot shows a web page titled "Gewissensbits" under the heading "Fallbeispiele zu Informatik und Ethik". The main image is a painting by Georges Seurat, depicting several figures in a park-like setting. On the left, there is a sidebar with navigation links: "GESELLSCHAFT FÜR INFORMATIK E.V.", "Fallbeispiele", "Chronologisch", "English Scenarios", "Aktuelles", "Kommentare", and "Debora Weber-Wulff bei". The main content area features the title "Fallbeispiel: Im seelsorgerischen KI-Gespräch" and author information "Debora Weber-Wulff, Constanze Kurz". The text discusses the lack of pastoral care in churches and how AI could be trained to provide it. It includes quotes from Matthias, Oktay, and Emma, and ends with a note from Matthias about the complexity of religious variations.

**Gewissensbits**  
Fallbeispiele zu Informatik und Ethik

Fallbeispiele Chronologisch English Scenarios Aktuelles Kommentare Debora Weber-Wulff bei

**Fallbeispiel: Im seelsorgerischen KI-Gespräch**

*Debora Weber-Wulff, Constanze Kurz*

In kirchlichen Gemeinden fehlt es heutzutage oft an Menschen, die Seelsorge leisten können. Kann eine KI hier die Lösung sein?

Im ländlichen Raum spitzt sich die Lage immer mehr zu: Es gibt kaum Pastorinnen oder Pastoren, die seelsorgerisch tätig sind. Das Start-up KI-Talks will daher eine KI mit einem besonderen Textkorpus trainieren, sodass diese in der Seelsorge eingesetzt werden kann.

Matthias, der Geschäftsführer von KI-Talks, ist bei einem Team-Meeting voller Enthusiasmus. „Wir können die schon vortrainierte KI LLaMA von Meta nehmen und sie mit bestimmten Texten weiter trainieren. Also nicht nur mit Texten aus der Bibel und viel theologischer Literatur – wir könnten auch gleich alle Predigten, die im Internet auf Deutsch zu finden sind, in die Trainingsdaten geben.“ Er denkt außerdem darüber nach, ob es sinnvoll wäre, ältere Bibelübersetzungen hinzuzunehmen.

Oktay, Senior Engineer bei KI-Talks, fragt in die Runde, warum es denn eigentlich nur christliche Texte sein sollen. Sollte nicht auch der Koran hinein? Und die vielen jüdischen Auslegungen der Tora? Er fragt auch: „Sollten wir nicht gleich mehrere verschiedene LLaMAs trainieren? So können die Leute auswählen, welches sie haben wollen. Vielleicht gibt es sogar einen Vergleichsmodus, wo man sehen kann, was der Imam meint, was der Rabbi oder was die Pastorin?“

Emma, Frontend Engineer bei KI-Talks, schnaubt: „Ihr glaubt wohl, es gibt nur eine Auslegung? Was ist mit feministischer Theologie? Und dann müssen wir auch die ganzen Fundamentalisten mit ihren wortgetreuen Bibelauslegungen beachten: Sollen die auch abgebildet werden? Und kann LLaMA die wirklich alle auseinanderhalten? Es gibt auch weitere Religionsunterschiede: Ich glaube, die Katholiken sehen die Welt ein bisschen anders als die Lutheraner.“

Matthias wirft ein, dass man die Varianten sehr wohl unterscheiden könnte. Ähnlich wie beim Segensroboter BlessU-2 [1], bei dem man die Art des Segens (Ermutigung, Erneuerung, Begleitung oder traditionell) auswählen kann, sollte man zum Beginn des Gesprächs einfach auf den gewünschten Religionsknopf klicken.

Emma meint, es sei überhaupt nicht klar, ob man verschiedene Religionsvarianten automatisch auseinanderhalten kann. Gerade bei der Unterscheidung von Fundamentalismus und weniger radikalen Auslegungen werde ein Knopfdruck nicht ausreichen, da brauche man eher einen Schieberegler. Auf jeden Fall müsse man erst einmal ausprobieren, ob so etwas überhaupt möglich sei.

→ Training erleichtert es, moralische Dilemmata zu erkennen

# Lesson 13: Technik-, Informations- und informatische Berufsethik & Nachhaltigkeit II

Eine spezifische „Informatik-Ethik“?

Möglichkeiten (und Grenzen) zur Einbindung in Entwicklungsprozesse

Ethisch motivierte Orientierungshilfen / Leitplanken

**Nachhaltigkeit als verwandte Herausforderung**

*„In den Studiengängen werden zum frühestmöglichen Zeitpunkt die Regeln guter wissenschaftlicher Praxis vermittelt und fortlaufend trainiert. Die Studierenden lernen, Wissen und Handeln in einen **übergeordneten** historischen, sozialen und kulturellen Kontext zu stellen und ethische Folgen des Handelns zu bedenken, um so zu einer **nachhaltigen Entwicklung** beitragen zu können.“*

§44(3) AllgStuPO

# Themenportal Klimaschutz

Forschung zum Nutzen der Gesellschaft – so lautet der Anspruch der TU Berlin. Forschung zu Klimaschutz und Nachhaltigkeit stehen dabei ganz oben auf der Agenda. Aber die TU Berlin tut mehr: Sie integriert das Thema in die Lehre, bietet spezielle Module und Studiengänge, setzt die Erkenntnisse ihrer Wissenschaftler\*innen auf dem Weg zu einem klimaneutralen Campus praktisch um und unterstützt ihre Mitglieder in deren persönlichem Engagement für den Klimaschutz.

Die Ergebnisse des Weltklimarats (IPCC) sind eindeutig: Der gegenwärtige Klimawandel ist Fakt und beruht vorwiegend auf menschlichen Einflüssen. Ebenso einig sind sich die Expert\*innen darin, dass längst nicht alle wissenschaftlichen Fragen beantwortet sind. Die weitere wissenschaftliche Erforschung des Themas hat eine hohe gesellschaftliche Priorität und damit auch eine hohe Priorität für die TU Berlin.

Die Klimakrise betrifft uns alle. Wir sind aufgefordert, aktiv zu werden, egal ob in Forschung und Lehre, in der Campusgestaltung oder in unserem individuellen Verhalten.

|  |   |   |  |
|--|---|---|--|
| Ansprechpartner*innen     | Unser Klima  | Nachhaltigkeitszertifikat  | Fridays for Future  |
| Gemeinsam sind wir dabei  | Unser Klima  |   |  |

## Aktuelles

### Krise der Demokratie im Huckepack der Klimakrise?

TU-Politikwissenschaftlerin Henrike Knappe über Zweifel, ob Demokratien das Klima wirklich schützen, über rechtspopulistisches Wahlverhalten und Klimaleugnung

24.01.2024

### Staffelstabübergabe beim Climate Change Center Berlin Brandenburg

## Ansprechpartner\*innen

-  Climate Action Task Force
-  Nachhaltigkeitsrat TU Berlin
-  Fridays for Future TU Berlin
-  Prof. Dr. Sophia Becker, Vizepräsidentin

<https://www.tu.berlin/themen/klimaschutz>

**Klimaschutzvereinbarung**  
zwischen dem  
Land Berlin  
**Senatsverwaltung für Umwelt, Verkehr und Klimaschutz**  
vertreten durch  
die Senatorin für Umwelt, Verkehr und Klimaschutz  
Frau Regine Günther  
  
und der  
**Technische Universität Berlin**  
vertreten durch  
den Präsidenten  
Herrn Prof. Dr. Christian Thomsen

## II. Ausgangssituation

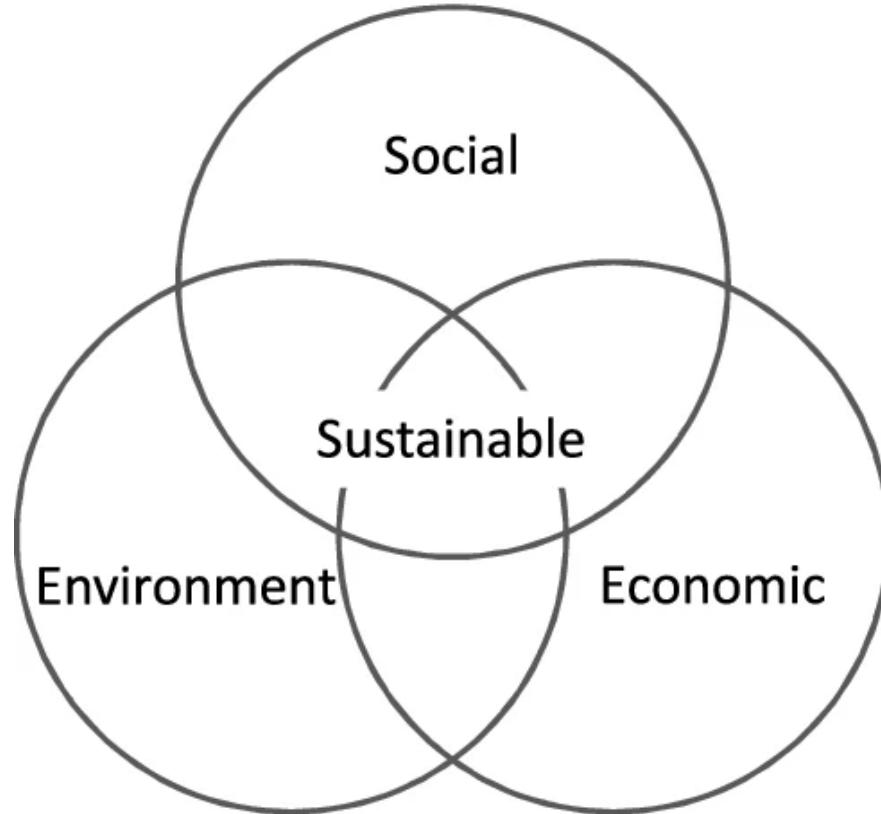
Angesichts der heutigen wissenschaftlichen Erkenntnisse besteht weitgehende Einigkeit darüber, dass der sparsame und effiziente Einsatz von Energie kurz- und mittelfristig die wichtigste Säule einer zukunftsfähigen und klimagerechten Energiepolitik darstellt. Entsprechend ambitioniert sind die Klimaschutzziele des Landes Berlin. Zur Erreichung dieser Ziele ist die Unterstützung aller Akteure der Stadtgesellschaft notwendig.

Die TU Berlin bekennt sich zu den unter § 3, Absatz 1 EWG genannten Klimaschutzzieilen und erklärt sich mit der vorliegenden Klimaschutzvereinbarung dazu bereit, das Land Berlin im Rahmen ihrer Möglichkeiten bei deren Erreichung zu unterstützen.

Die Grundlage für die vorliegende Vereinbarung bildet der Gebäudebezogene Energieverbrauch des Basisjahres 2018 (siehe Anlage 1). Der damit verbundene CO<sub>2</sub>-Ausstoß<sup>1</sup>, der als Basis für das unter Kapitel III vereinbarte Einsparziel dient, betrug 45.934 Tonnen. Das entspricht bei einer NettoGESCHOSSENFÄLCE (NGF) von 627.185 m<sup>2</sup> einem Wert von 73,2 kg CO<sub>2</sub>/m<sup>2</sup>NGF. Der Energieverbrauch wird hauptsächlich verursacht durch die

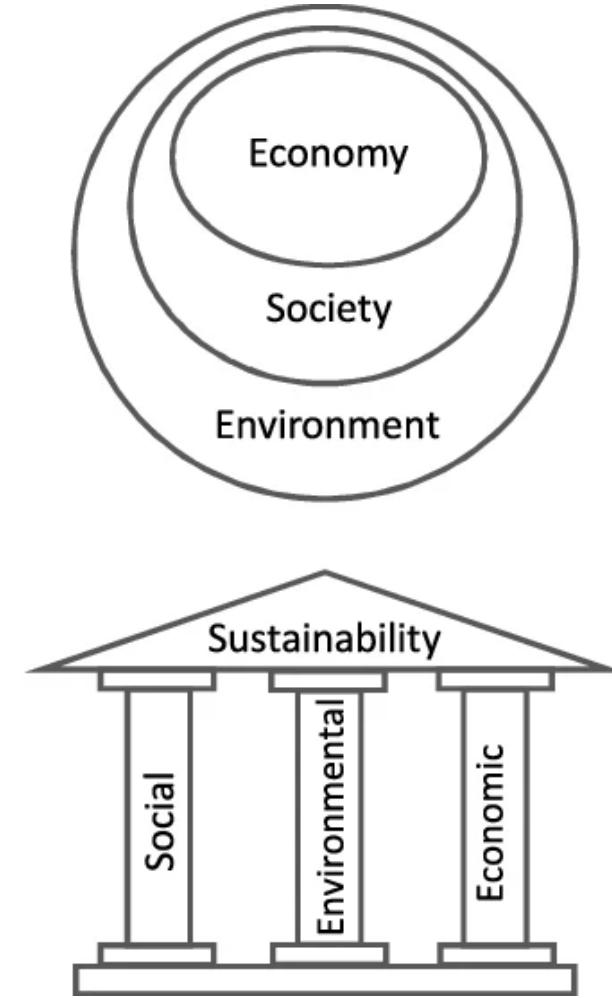
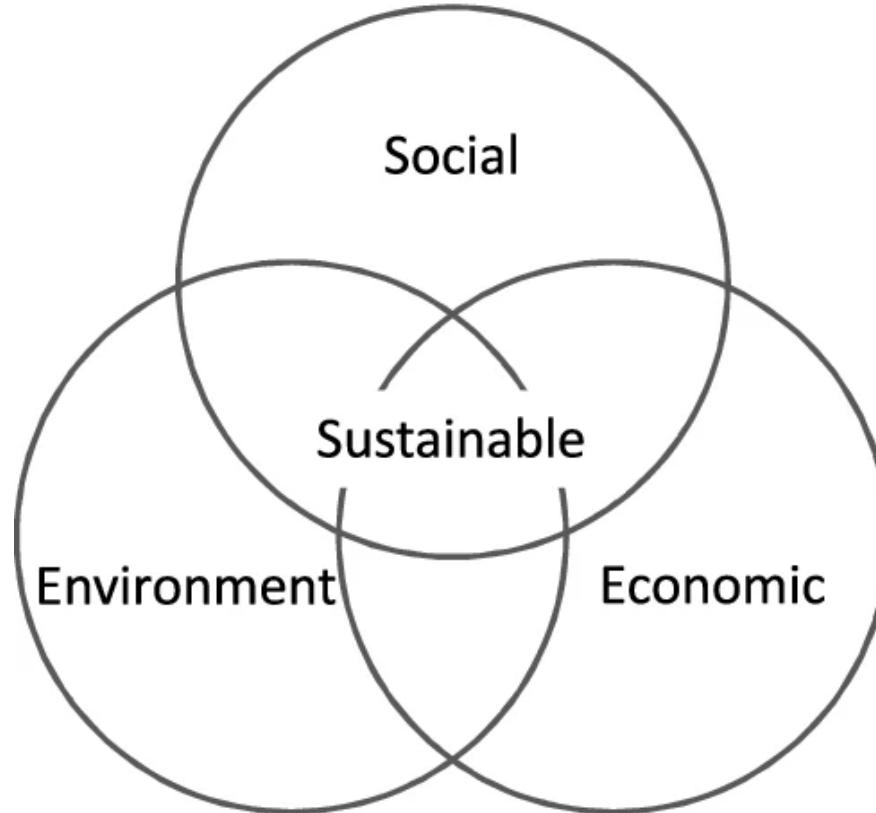
<sup>1</sup> Zur Ermittlung der energieverbrauchsbedingten CO<sub>2</sub>-Emissionen werden die vom Amt für Statistik in der offiziellen Energie- und CO<sub>2</sub>-Bilanz für das Jahr 2018 veröffentlichten Emissionsfaktoren verwendet.

# „Traditionelle“ Nachhaltigkeit



<https://en.wikipedia.org/wiki/Sustainability>

# „Traditionelle“ Nachhaltigkeit



<https://en.wikipedia.org/wiki/Sustainability>

# Grundproblem

Atmosphäre als  
begrenzter „Deponieraum“

1100 GtCO<sub>2</sub> für 2°C-Ziel  
400 GtCO<sub>2</sub> für 1,5°C-Ziel

## Die CO<sub>2</sub>-Uhr tickt

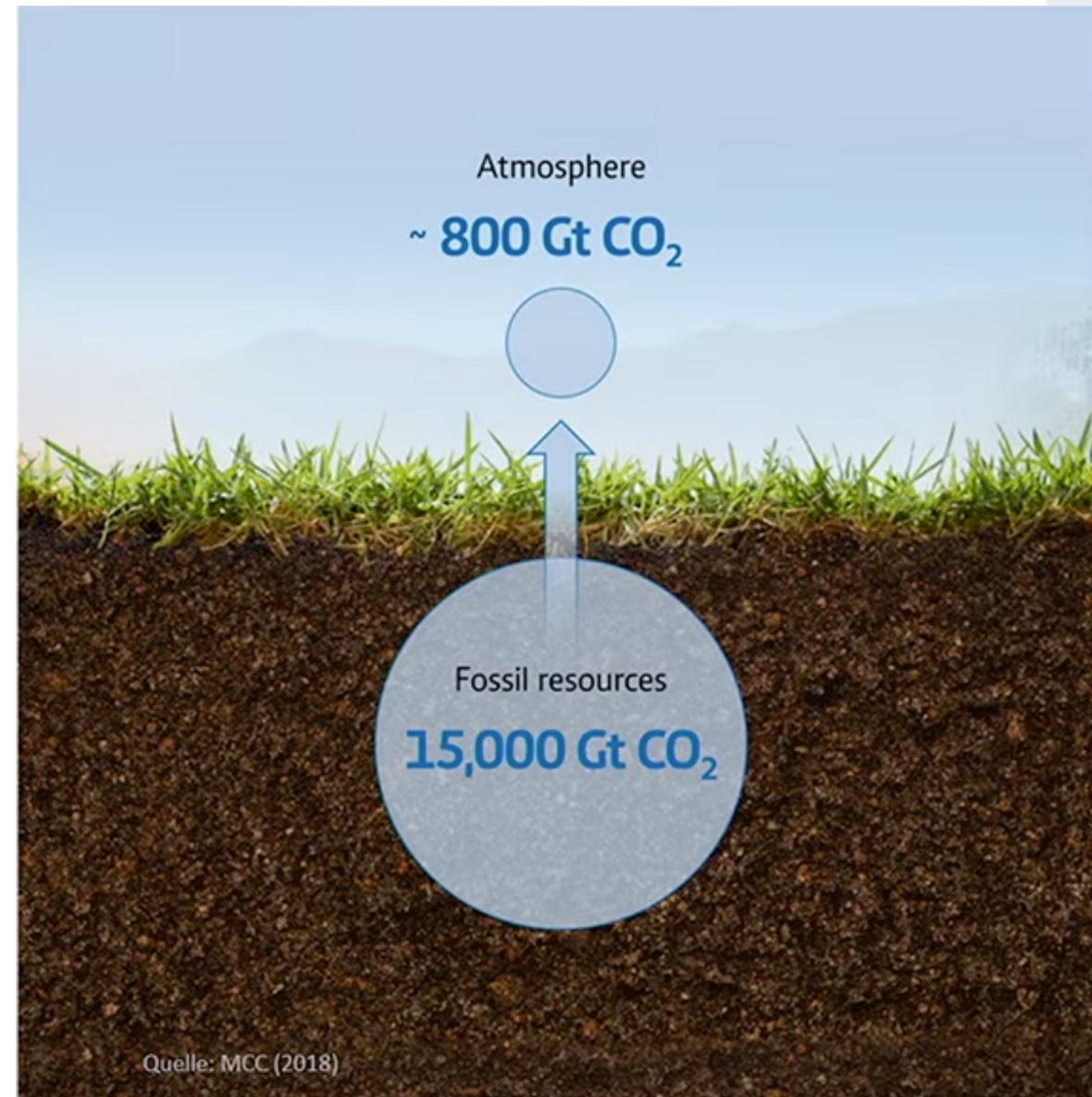
Zeit zur Einhaltung der 2-Grad-Grenze

| Jahr | Monat | Tag | Std. | Min. | Sek.  |
|------|-------|-----|------|------|-------|
| 23   | 3     | 2   | 6    | 52   | 37 17 |

Verbleibendes CO<sub>2</sub>-Budget

981'286'609'337

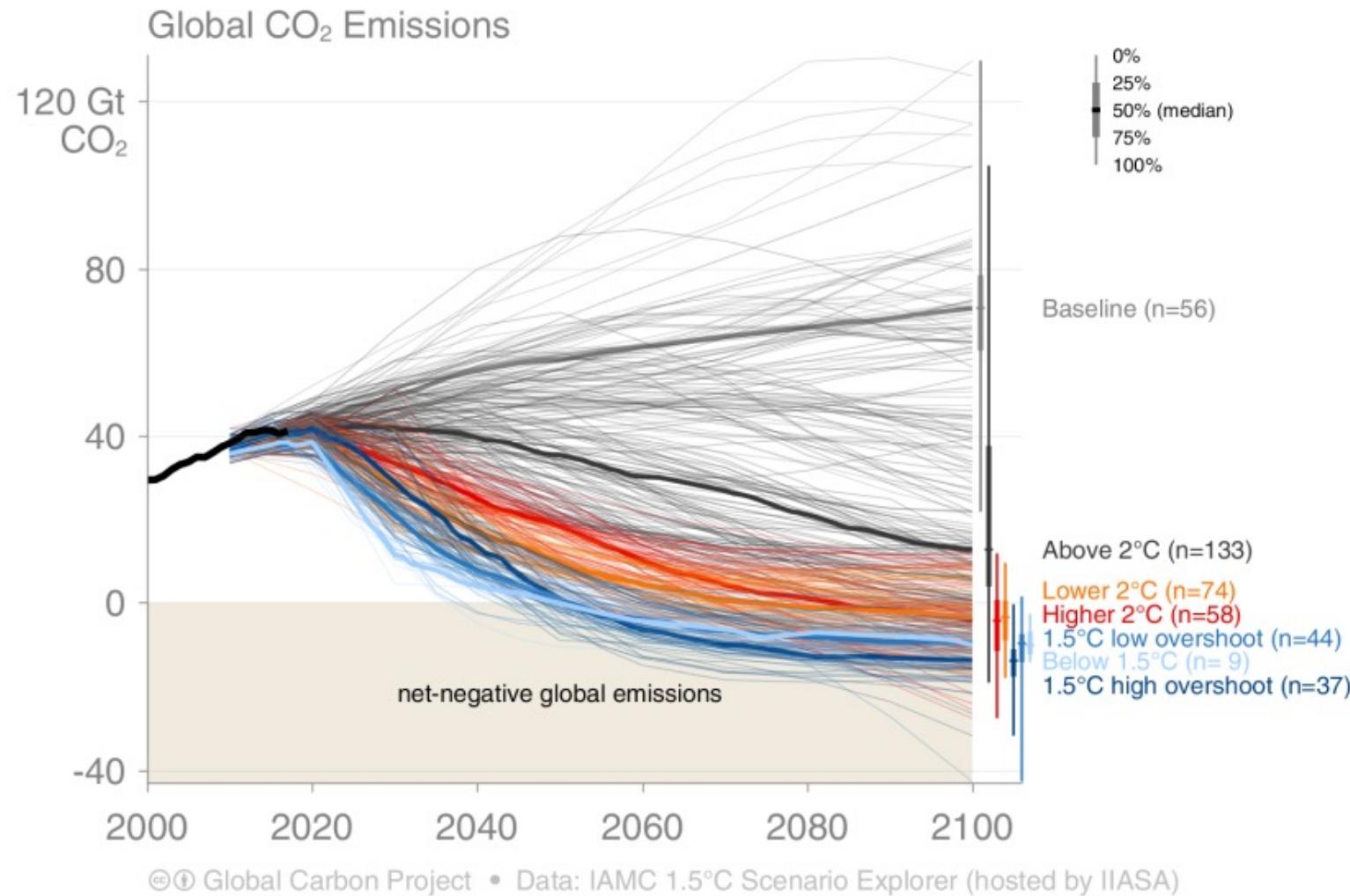
<https://www.mcc-berlin.net/forschung/co2-budget.html>



Edenhofer 2020

# Emissionen

„CO<sub>2</sub>-Emissionen zur Einhaltung der 1.5- bzw. 2.0-Grad-Grenze bis 2100. Links sind die jährlichen CO<sub>2</sub>-Emissionen angegeben. Die dünnen Linien zeigen unterschiedliche Modellberechnungen, die dicken Linien Mittelwerte. Zur Einhaltung der Ziele von Paris sind in fast allen Fällen negative Emissionen notwendig.“



<https://wiki.bildungsserver.de/klimawandel/index.php/2-Grad-Ziel>

# Wo stehen wir?



CO2-Uhr springt auf drei Jahre

## Der 1,5-Grad-Countdown

Die Klima-Uhr der taz zeigt: Nur noch drei Jahre, dann ist das weltweite CO2-Budget für 1,5 Grad abgelaufen. Rasches Handeln wird immer dringlicher.

TAZ, 01.10.2023, <https://taz.de/CO2-Uhr-springt-auf-drei-Jahre/!5961005/>

Wetterrückblick 2023

## Das heißeste Jahr aller Zeiten

ZEIT, 09.01.2024

Umweltbundesamt

## Folgen der Klimakrise in Deutschland verschärfen sich

Deutschland gehört zu den Regionen mit dem höchsten Wasserverlust weltweit. Wegen der klimabedingten andauernden Trockenheit und des damit...

28.11.2023

UBA et al.

Pallas / Grünwald – Information Governance

53



Die CO<sub>2</sub>-Uhr tickt

2 7 10 6 0 17

Jahre Monate Tage Stunden Minuten Sekunden

bleiben, bis das globale CO<sub>2</sub>-Budget für das Erreichen des 1,5-Grad-Limits aufgebraucht ist.



► Alles zur Klimakrise



VIELE OPFER, HOHE SCHÄDEN

Der Klimawandel ist da: Wir müssen handeln

Der Klimawandel ist da, auch in Deutschland. Jetzt muss alles getan werden, um wenigstens die Folgen abzumildern.

Archibald Preuschat

15.01.2024, 15:21 Uhr

FAZ, 05.01.2024

SEengineering

## „Zwischen Klimademos und Klimakabinett – wie die CO<sub>2</sub>-Bepreisung gelingen kann“

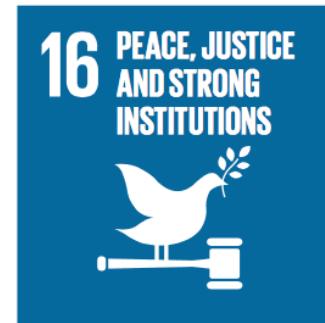
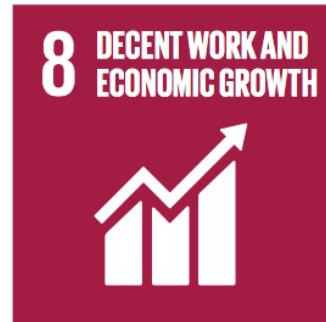
Prof. Dr. Ottmar Edenhofer

Direktor des Potsdam-Instituts für Klimafolgenforschung und  
Professor für Ökonomie des Klimawandels an der TU Berlin

Gibt es eine ethische Verpflichtung  
treibhausgasneutral zu leben?

- Niemand kann auf individueller Ebene treibhausgasneutral leben.
- Ein **antizipatorischer** Lebensstil kann ein Zeichen setzen: „Ich lebe so, als ob wir schon in einer treibhausgasneutralen Gesellschaft lebten“ (weniger Fleisch, weniger Flugreisen).
- Eine verändernde Kraft wird dieser Lebensstil nur dann entfalten, wenn wir zugleich von der **Politik** die Änderung der Rahmenbedingungen fordern: Einen CO<sub>2</sub>-Preis!





# UN Sustainable Development Goals

*“Halfway to the deadline for the 2030 Agenda, the SDG Progress Report; Special Edition shows we are leaving more than half the world behind. Progress on more than 50 per cent of targets of the SDGs is weak and insufficient; on 30 per cent, it has stalled or gone into reverse. These include key targets on poverty, hunger and climate. Unless we act now, the 2030 Agenda could become an epitaph for a world that might have been.” – António Guterres (Secretary-General of the United Nations)*



## The Sustainable Development Goals Report 2023: Special edition

Towards a Rescue Plan for People and Planet



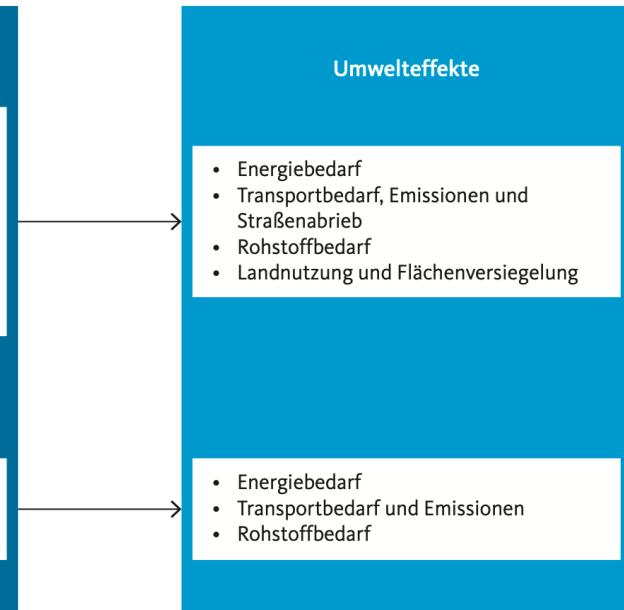
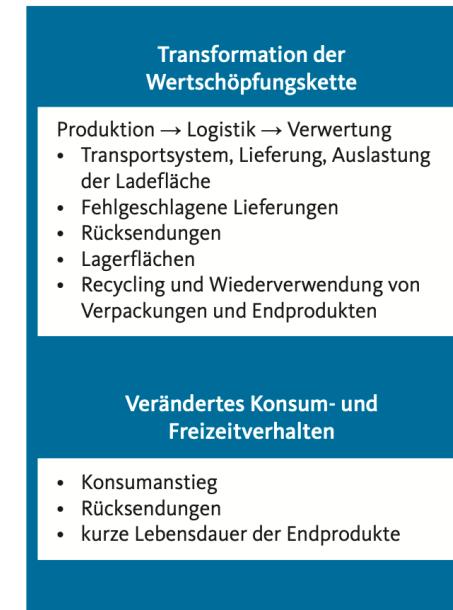


## Inhaltsverzeichnis

|         |  |     |
|---------|--|-----|
| 5.2.1.3 | Effizientere Produktionsprozesse durch Industrie 4.0 und digital koordinierte Kreislaufwirtschaft .....            | 161 |
| 5.2.1.4 | Folgerungen: Was kann Digitalisierung für die globale Transformation des industriellen Metabolismus leisten? ..... | 163 |
| 5.2.2   | Neue Formen digitaler Ökonomie: Ansätze des nachhaltigen Wirtschaftens im Digitalen Zeitalter .....                | 165 |
| 5.2.2.1 | Wiedereinbettung als Herausforderung nachhaltigen Wirtschaftens.....   | 165 |
| 5.2.2.2 | Nachhaltiges digitales Unternehmertum .....  | 166 |
| 5.2.2.3 | Plattformkooperativen als besonderer Ausdruck kollektiven Unternehmertums.....                                     | 166 |
| 5.2.2.4 | Sharing-Ökonomie zwischen klassischer und kollektiver Ökonomie .....   | 167 |
| 5.2.2.5 | Kollaborative Produktionsformen: Prosumenten und Commons-based Peer Production .....                               | 170 |
| 5.2.2.6 | Folgerungen .....  | 171 |
| 5.2.3   | Digitalisierung des Konsums und nachhaltiges Konsumverhalten: Förderung solidarischer Lebensstile .....            | 171 |
| 5.2.3.1 | Digitalisierter Konsum zur Erhaltung natürlicher Lebensgrundlagen .....  | 171 |
| 5.2.3.2 | Von der Erhaltung der natürlichen Lebensgrundlagen zum Konzept der solidarischen Lebensqualität .....              | 175 |
| 5.2.3.3 | Chancen und Risiken des digitalisierten Konsums für Teilhabe und Eigenart .....                                    | 175 |
|         | <i>Themenkasten 5.2-1 FinTech im Kontext nachhaltiger Finanzierung .....</i>                                       | 176 |
| 5.2.4   | Nachhaltigkeit beim Onlinehandel: Status Quo und Perspektiven .....  | 178 |
| 5.2.4.1 | Rolle und Wachstum des Onlinehandels .....   | 178 |
| 5.2.4.2 | Umweltauswirkungen des Onlinehandels .....   | 179 |
| 5.2.4.3 | Soziale Effekte .....  | 181 |
| 5.2.4.4 | Folgerungen .....  | 183 |
| 5.2.5   | Digitalisierung: vom Elektroschrottproblem zur Lösung für Kreislaufwirtschaft? .....                               | 184 |
| 5.2.5.1 | Elektroschrott im Kontext der Kreislaufwirtschaft .....  | 184 |
| 5.2.5.2 | Digitale Technologien als Ursache des globalen Elektroschrottproblems .....  | 184 |
| 5.2.5.3 | Digitale Technologien zur Lösung des Elektroschrottproblems .....  | 185 |
| 5.2.5.4 | Folgerungen .....  | 189 |
| 5.2.6   | Digitalisierung für Klimaschutz und Energiewende .....   | 190 |
| 5.2.6.1 | Digitale Technologien für die Energiewende nutzen .....  | 191 |
| 5.2.6.2 | Digitale Technologien zur Überwindung von Energiearmut in Entwicklungsländern nutzen .....                         | 192 |
| 5.2.6.3 | Durch Digitalisierung erzeugte Energienachfrage einhegen.....  | 193 |
| 5.2.6.4 | Risiken eines digitalisierten Energiesystems: Resilienz und Privatsphäre ..  | 194 |
| 5.2.6.5 | Folgerungen .....  | 196 |
| 5.2.7   | Smart City: Nachhaltige Stadtentwicklung mit Digitalisierung? .....  | 197 |
| 5.2.7.1 | Nachhaltige Stadtentwicklung im Digitalen Zeitalter: Herausforderungen   | 197 |
| 5.2.7.2 | Smart City: Konzept, Anwendungsbeispiele, Verbreitung und Treiber .....  | 197 |
| 5.2.7.3 | Ausgewählte Spannungsfelder digital unterstützter Stadtentwicklung .....   | 199 |
| 5.2.7.4 | Digitale Souveränität und das „Recht auf Stadt“ .....  | 200 |
| 5.2.7.5 | Folgerungen .....  | 202 |
| 5.2.8   | Nachhaltige urbane Mobilität im Digitalen Zeitalter .....  | 204 |
| 5.2.8.1 | Leitbilder einer nachhaltigen urbanen Mobilitätswende .....  | 204 |
| 5.2.8.2 | Elemente der digitalen Mobilitätswende .....   | 204 |
| 5.2.8.3 | Status Quo und Herausforderungen der nachhaltigen digitalen Mobilität im urbanen Raum .....                        | 207 |
| 5.2.8.4 | Folgerungen: Stellschrauben für eine nachhaltige digitale Mobilitätswende im urbanen Raum .....                    | 208 |
| 5.2.9   | Präzisionslandwirtschaft: der nächste Schritt in die industrialisierte Landwirtschaft?.....                        | 210 |

[https://www.wbgu.de/de/publikationen/publikation/unser-gemeinsame-digitale-zukunft-\(2019\)](https://www.wbgu.de/de/publikationen/publikation/unser-gemeinsame-digitale-zukunft-(2019))

# Beispiel: Online-Handel



**Abbildung 5.2.4-1**

Umwelteffekte des B2C-Onlinehandels.

Quelle: WBGU in Anlehnung an Fichter (2003) und Tiwari und Singh (2011)

# Beispiel: Elektroschrott

Titel | Risiko Cloudabschaltung c't 3/2024 S. 18



Bild: Moritz Reichartz

## Aus allen Wolken gefallen

**Risiko Cloudabschaltung:** Wenn einwandfreie Hardware plötzlich nutzlos wird

Von der Überwachungskamera über die Smart-Home-Zentrale bis zum E-Bike: Schaltet der Hersteller die Clouddienste ab, wird teure Hardware über Nacht zu Elektroschrott. Das Problem wird sich weiter verschärfen, denn die Cloudifizierung schreitet voran – und die Rechtslage schützt Nutzer nicht vor der Willkür der Hersteller.

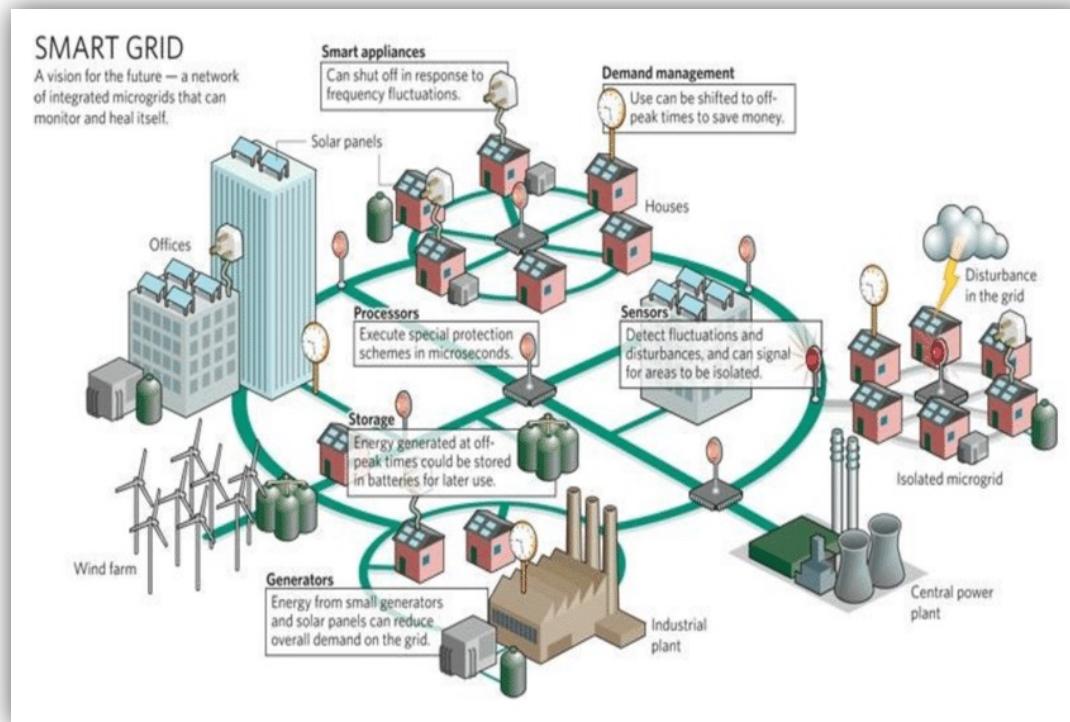
**Modell der Kreislaufwirtschaft:**  
weniger Rohstoffe, weniger Abfall, weniger Emissionen



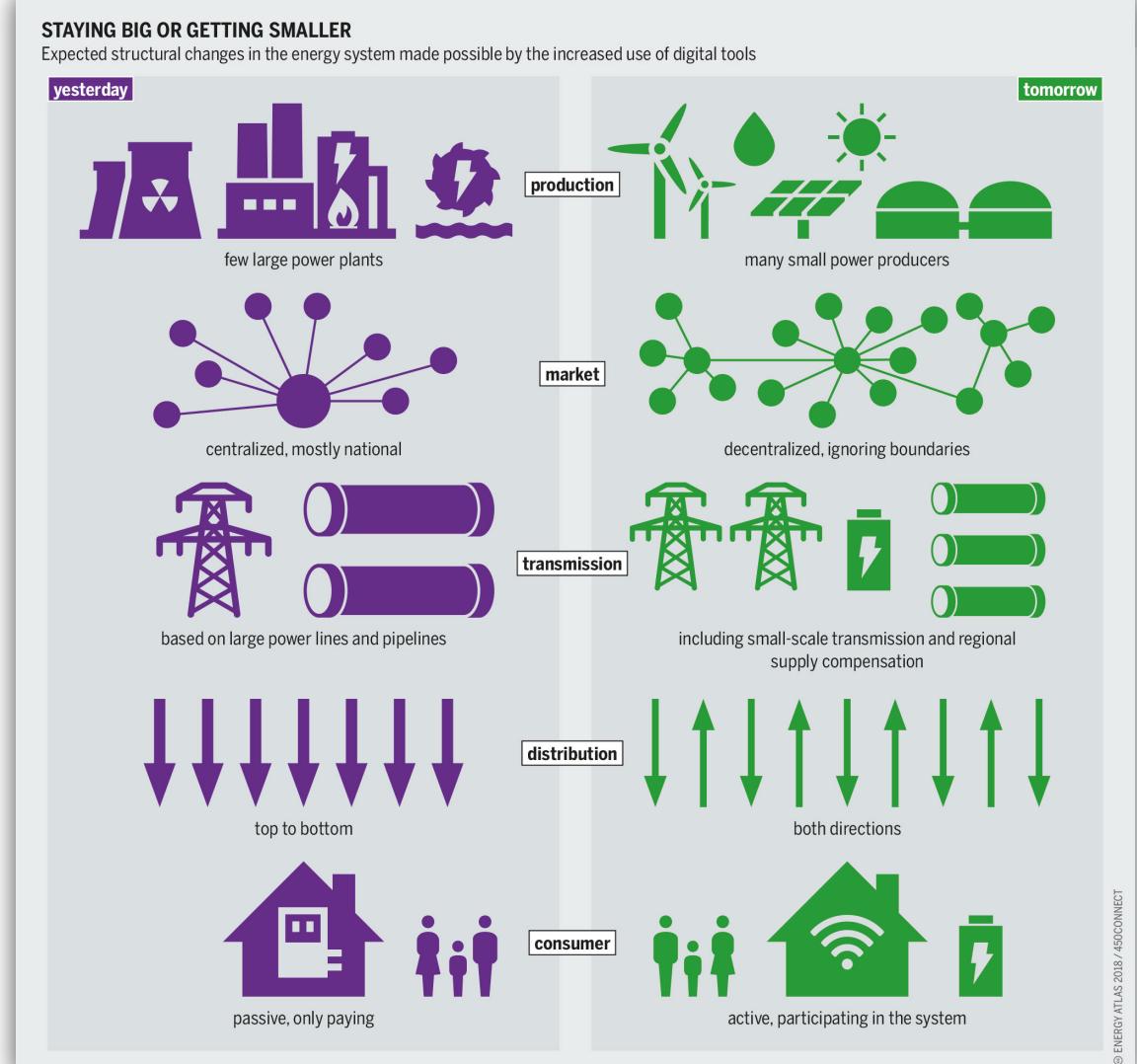
Quelle: Wissenschaftlicher Dienst des Europäischen Parlaments



# Beispiel: Energieversorgung / Smart Grid



[https://miis.maths.ox.ac.uk/732/1/edf\\_draft\\_FixedConcurrentGenCos\\_vboris.pdf](https://miis.maths.ox.ac.uk/732/1/edf_draft_FixedConcurrentGenCos_vboris.pdf)

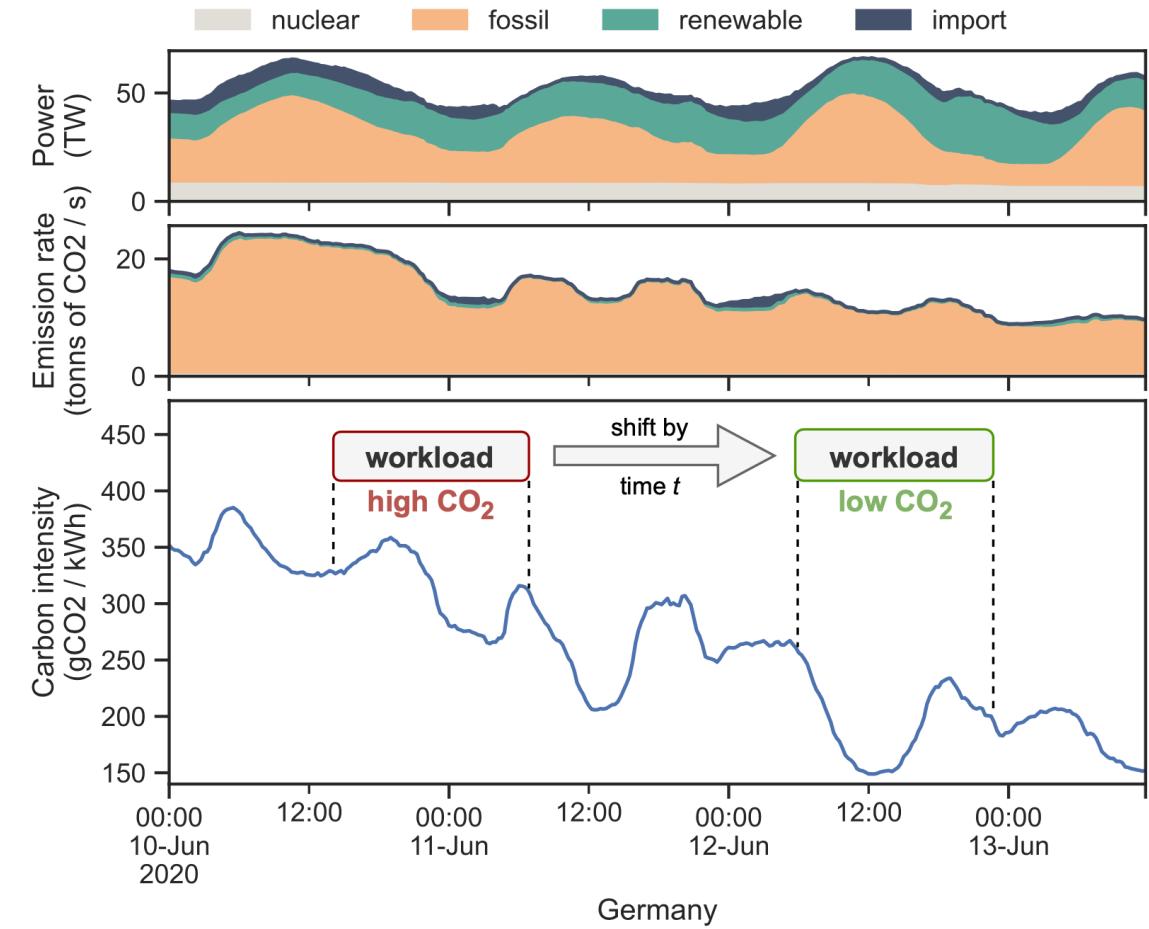


# Beispiel: Cloud Computing / AI training workloads

**"GPT-3 with 175 billion parameters, emitted over 550 tons of CO<sub>2</sub><sub>e</sub>** while consuming 1,287 MW hours of electricity, per computer scientist Kate Saenko. It's the same amount of emissions as a single person taking **550 roundtrip flights between New York and San Francisco.**

And that doesn't even include other sources of emissions, only getting the AI ready to use."

<https://carboncredits.com/how-big-is-the-co2-footprint-of-ai-models-chatgpts-emissions/>



<https://dl.acm.org/doi/10.1145/3464298.3493399>

# Beispiel: Large-language models

“increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption”

“the tendency of human interlocutors to impute meaning where there is none can mislead both NLP researchers and the general public into taking synthetic text as meaningful”

“Size Doesn’t Guarantee Diversity”

“Encoding Bias”

Stochastic parrot: “haphazardly stitching together sequences of linguistic forms ... according to probabilistic information about how they combine, but without any reference to meaning” (Weil/Bender 2021)

## On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender\*  
ebender@uw.edu  
University of Washington  
Seattle, WA, USA

Angelina McMillan-Major  
aymm@uw.edu  
University of Washington  
Seattle, WA, USA

Timnit Gebru\*  
timnit@blackinai.org  
Black in AI  
Palo Alto, CA, USA

Shmargaret Shmitchell  
shmargaret.shmitchell@gmail.com  
The Aether

### ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

### CCS CONCEPTS

• Computing methodologies → Natural language processing

### ACM Reference Format:

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3442188.3445922>

### 1 INTRODUCTION

One of the biggest trends in natural language processing (NLP) has been the increasing size of language models (LMs) as measured by the number of parameters and size of training data. Since 2018

Joint first authors



This work is licensed under a Creative Commons Attribution International 4.0 License.  
FAccT '21, March 3–10, 2021, Virtual Event, Canada  
ACM ISBN 978-1-4503-8309-7/21/03.  
<https://doi.org/10.1145/3442188.3445922>

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

Just as environmental impact scales with model size, so does the difficulty of understanding what is in the training data. In §4, we discuss how large datasets based on texts from the Internet overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations. In collecting ever larger datasets we risk incurring documentation debt. We recommend mitigating these risks by budgeting for curation and documentation at the start of a project and only creating datasets as large as can be sufficiently documented.

As argued by Bender and Koller [14], it is important to understand the limitations of LMs and put their success in context. This not only helps reduce hype which can mislead the public and researchers themselves regarding the capabilities of these LMs, but might encourage new research directions that do not necessarily depend on having larger LMs. As we discuss in §5, LMs are not performing natural language understanding (NLU), and only have success in tasks that can be approached by manipulating linguistic form [14]. Focusing on state-of-the-art results on leaderboards without encouraging deeper understanding of the mechanism by which they are achieved can cause misleading results as shown

# Beispiel: Large-language models

## *Google Researcher Says She Was Fired Over Paper Highlighting Bias in A.I.*

Timnit Gebru, one of the few Black women in her field, had voiced exasperation over the company's response to efforts to increase minority hiring.

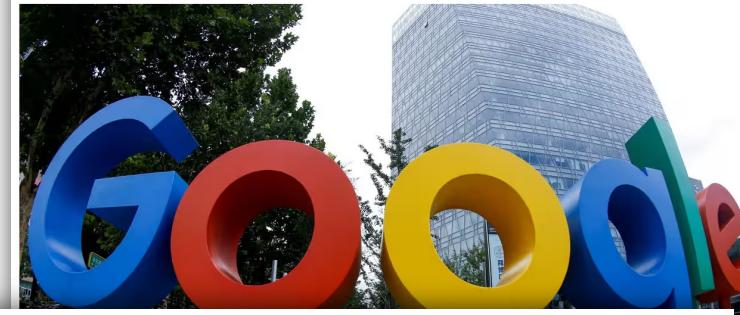
Share full article 276



Timnit Gebru, a respected researcher at Google, questioned biases built into artificial intelligence systems. Cody O'Loughlin for The New York Times

## Google fires Margaret Mitchell, another top researcher on its AI ethics team

The dismissal comes after prominent Black researcher Timnit Gebru was fired in December; both had called for more diversity among research staff



## More than 1,200 Google workers condemn firing of AI scientist Timnit Gebru

More than 1,500 researchers also sign letter after Black expert on ethics says Google tried to suppress her research on bias



## On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender\*  
ebender@uw.edu  
University of Washington  
Seattle, WA, USA

Angelina McMillan-Major  
aymm@uw.edu  
University of Washington  
Seattle, WA, USA

Timnit Gebru\*  
timni@blackinai.org  
Black in AI  
Palo Alto, CA, USA

Shmargaret Shmitchell  
shmargaret.shmitchell@gmail.com  
The Aether

### ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

### CCS CONCEPTS

• Computing methodologies → Natural language processing

#### ACM Reference Format:

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3442188.3445922>

### 1 INTRODUCTION

One of the biggest trends in natural language processing (NLP) has been the increasing size of language models (LMs) as measured by the number of parameters and size of training data. Since 2018

Joint first authors



This work is licensed under a Creative Commons Attribution International 4.0 License.  
*FAccT '21, March 3–10, 2021, Virtual Event, Canada*  
ACM ISBN 978-1-4503-8309-7/21/03.  
<https://doi.org/10.1145/3442188.3445922>

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (\$2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

Just as environmental impact scales with model size, so does the difficulty of understanding what is in the training data. In §4, we discuss how large datasets based on texts from the Internet overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations. In collecting ever larger datasets we risk incurring documentation debt. We recommend mitigating these risks by budgeting for curation and documentation at the start of a project and only creating datasets as large as can be sufficiently documented.

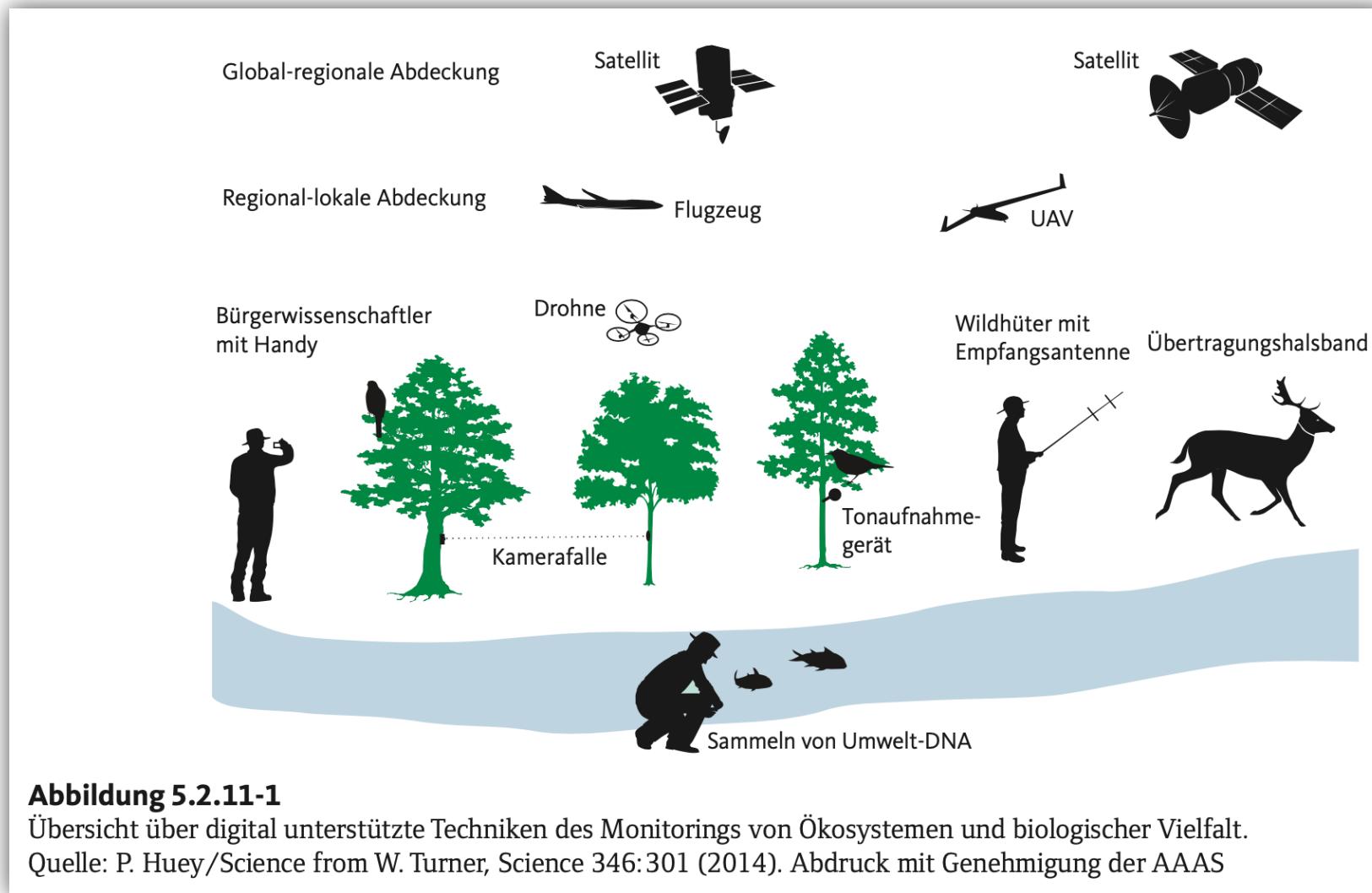
As argued by Bender and Koller [14], it is important to understand the limitations of LMs and put their success in context. This not only helps reduce hype which can mislead the public and researchers themselves regarding the capabilities of these LMs, but might encourage new research directions that do not necessarily depend on having larger LMs. As we discuss in §5, LMs are not performing natural language understanding (NLU), and only have success in tasks that can be approached by manipulating linguistic form [14]. Focusing on state-of-the-art results on leaderboards without encouraging deeper understanding of the mechanism by which they are achieved can cause misleading results as shown

# Beispiel: Smart mobility



<https://www.magility.com/smart-mobility-wohin-geht-die-reise/>

# Beispiel: Artenschutz / Biodiversität



**Abbildung 5.2.11-1**

Übersicht über digital unterstützte Techniken des Monitorings von Ökosystemen und biologischer Vielfalt.  
Quelle: P. Huey/Science from W. Turner, Science 346: 301 (2014). Abdruck mit Genehmigung der AAAS



*Ihr Job & private Lebensentscheidungen (individuelle Lebensführung), aber vor allem auch  
Ihr Wahlverhalten, politische Teilhabe... (kollektive Einflussfaktoren)*

Graue Energie

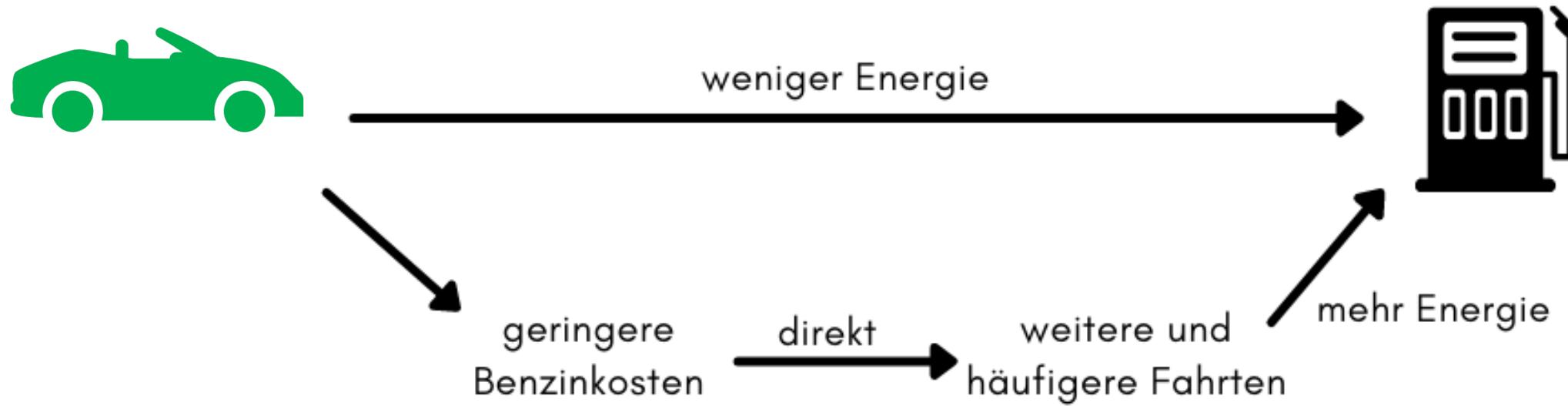


weniger Energie

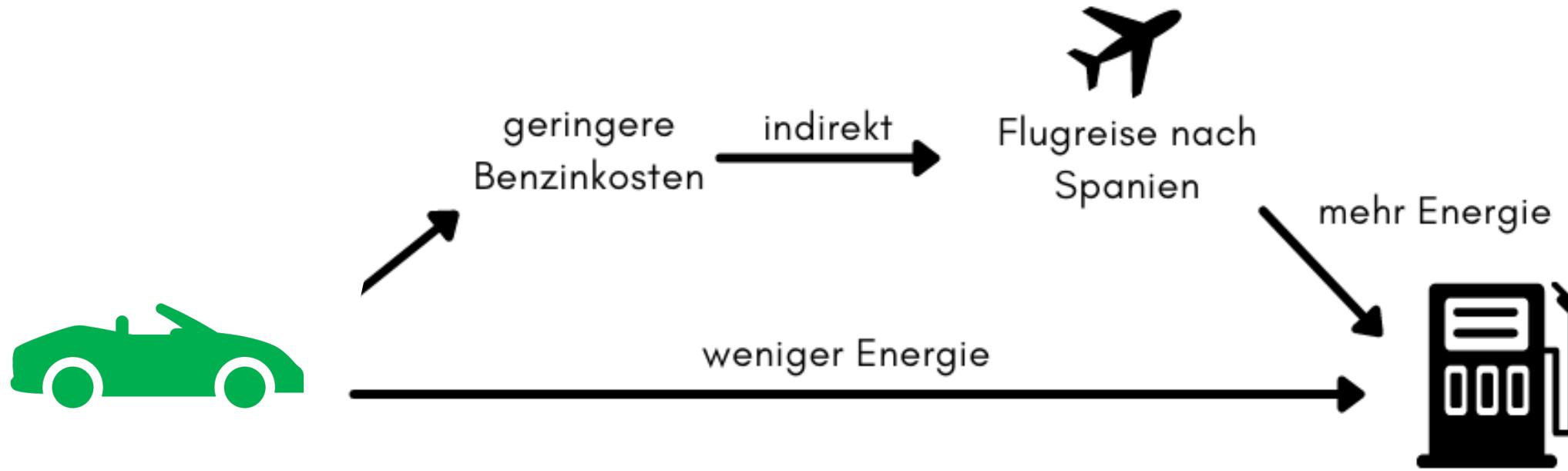




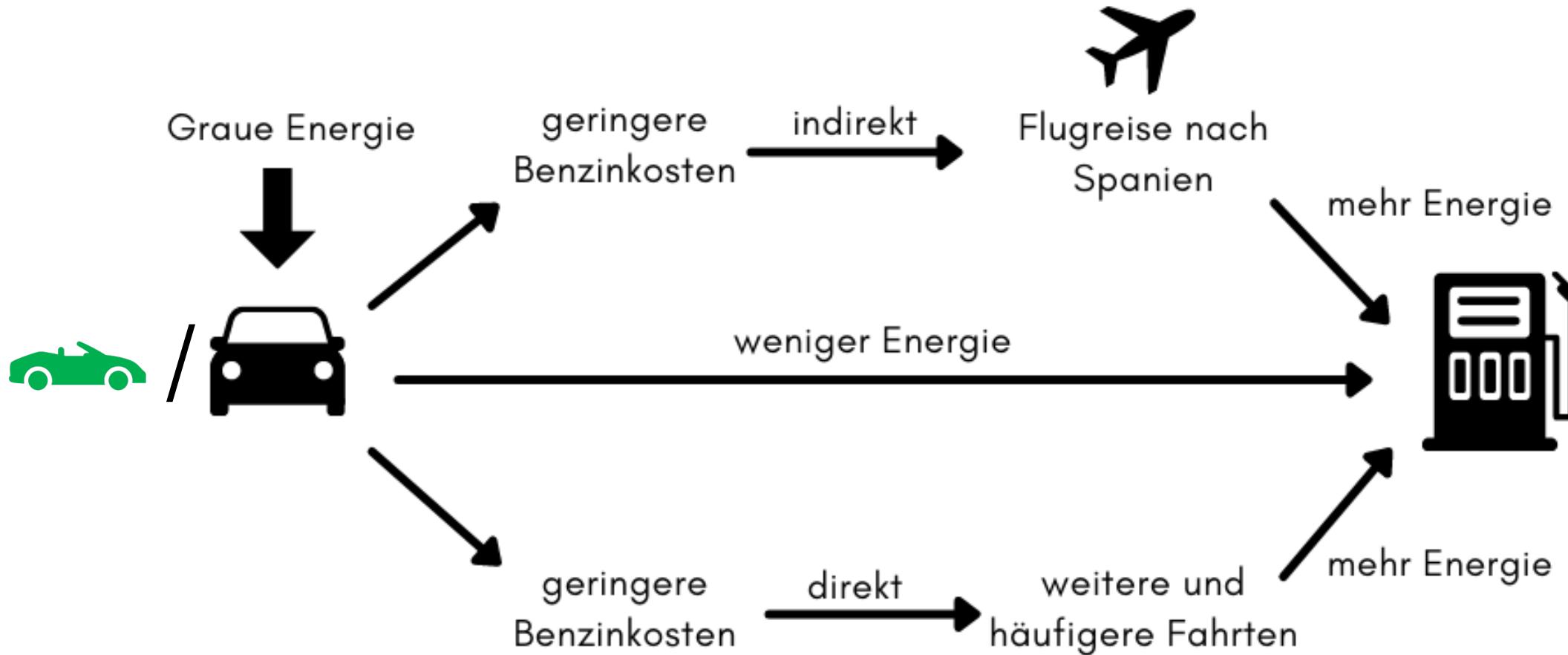
# Direkter Rebound-Effekt



# Indirekter Rebound-Effekt



# Rebound-Effekt



<https://kaufnix.net/arten-von-rebound-effekten/>

# Rebound-Effekt

- Differenz zwischen der theoretisch zu erwartenden Einsparung durch eine Effizienzmaßnahme und der tatsächlichen erreichten Einsparung
  - Je größer die Differenz zwischen der zuvor erwarteten und der tatsächlichen Einsparung, umso größer ist der Rebound-Effekt
- Effizienzsteigerungen sollten demnach so angelegt werden, dass echte Einsparung erzielt werden kann



Vgl. Umweltbundesamt

# Klimaschonende Maßnahmen

## Auswirkungen 1. Ordnung (1<sup>st</sup> Order Impact)

Resourcenschonung bei neuer Infrastruktur  
(insbesondere Hardware, Netzwerke etc.)

## Auswirkungen 2. Ordnung

Effizienz und Konsumeffekte durch neu-organisierte Prozesse

## Auswirkungen 3. Ordnung

Konsum- und Lifestyle-Veränderungen in der Gesellschaft

nach Maja Göpel, 2023, [https://media.ccc.de/v/37c3-12324-on\\_digitalisation\\_sustainability\\_climate\\_justice](https://media.ccc.de/v/37c3-12324-on_digitalisation_sustainability_climate_justice)

# Klimaschonende Maßnahmen

Potenziale der Digitalisierung für die **Minderung von Treibhausgasemissionen**

Auswirkungen von **digitalen Vermarktungsstrategien** auf das Konsumverhalten

**Gemeinwohlorientierung** im Zeitalter der Digitalisierung

nach UBA, <https://www.umweltbundesamt.de/themen/digitalisierung/digitale-nachhaltigkeit>

Technik- und informatische Berufsethik,  
Nachhaltigkeit I + II

## Potenzielle der Digitalisierung für die **Minderung von Treibhausgasemissionen**

Recht, Vertrag, Eigentum,  
Transaktion

## Auswirkungen von **digitalen Vermarktungsstrategien** auf das Konsumverhalten

Netzwerk- und Plattform-Ökon.,  
Digital Commons

Digitale Transformation

## **Gemeinwohlorientierung** im Zeitalter der Digitalisierung

Privacy Engineering

Technikbasierte Regulierung,  
„Code as Law“

nach UBA, <https://www.umweltbundesamt.de/themen/digitalisierung/digitale-nachhaltigkeit>

## Technik- und informatische Berufsethik, Nachhaltigkeit I + II

# Potenzielle der Digitalisierung für die **Minderung von Treibhausgasemissionen**

Recht, Vertrag, Eigentum,  
Transaktion

*Umweltexternalitäten in Tx, CO<sub>2</sub>-Preis,  
Klimaschutzgesetze,..*

## Auswirkungen von **digitalen Vermarktungsstrategien** auf das Konsumverhalten

Netzwerk- und Plattform-Ökon.,  
Digital Commons

Digitale Transformation

*Klimafolgen von digitalen  
Geschäftsmodellen & Technologien*

## **Gemeinwohlorientierung** im Zeitalter der Digitalisierung

Privacy Engineering

Technikbasierte Regulierung,  
„Code as Law“

*Öffentlich-rechtliche Infrastruktur,  
Soziale Standards*

nach UBA, <https://www.umweltbundesamt.de/themen/digitalisierung/digitale-nachhaltigkeit>

# Digitalisierung und Nachhaltigkeit

## Nachhaltige digitale Infrastrukturen & digitale Tools, z.B.

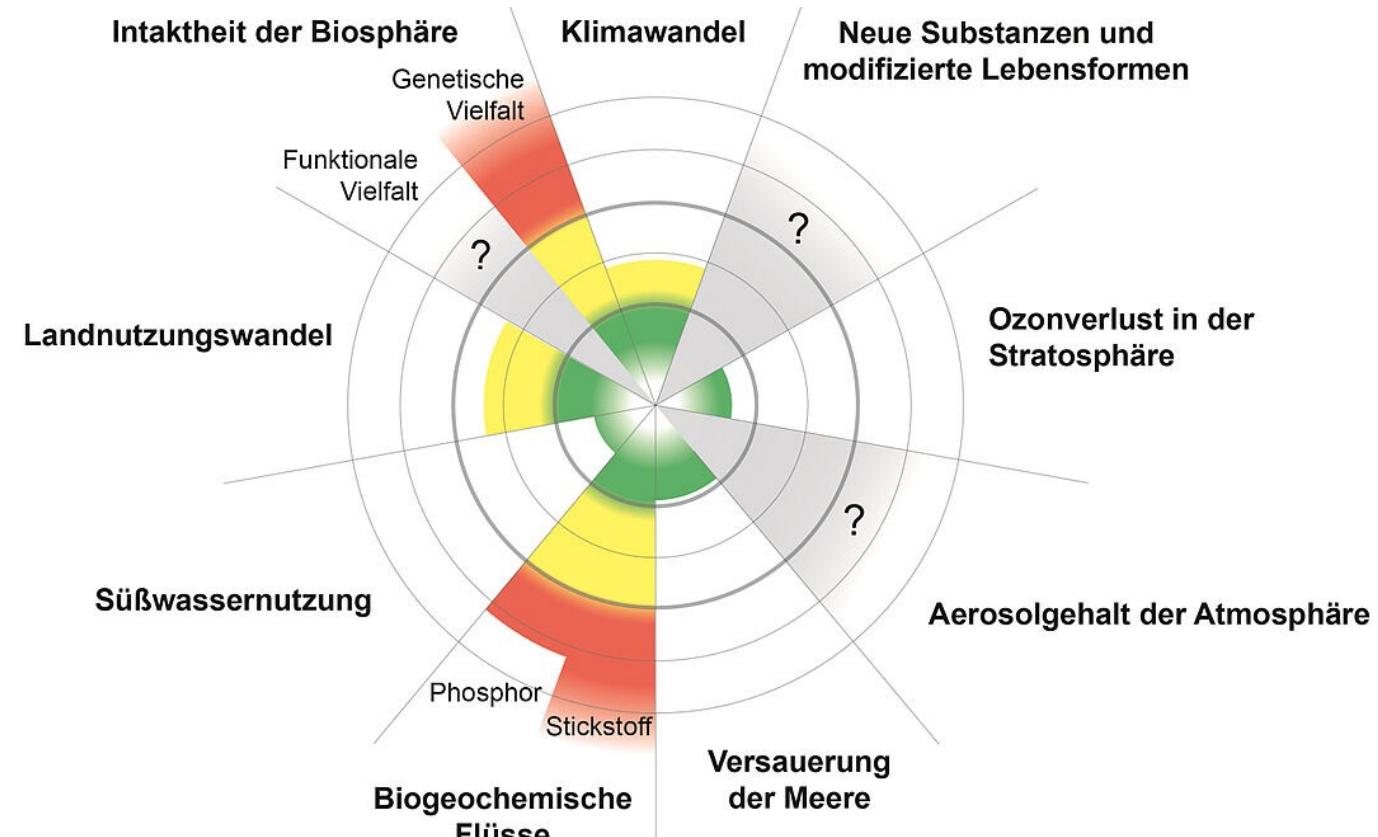
- Energie,
- Mobilität,
- Materialien,
- Entwicklungshilfe,...

## „Digital-nachhaltige“ Gesellschaften, z.B.

- sozialer Zusammenhalt (Digital Divide, Fairness...),
- Machtverhältnisse („Big Tech“,...),
- Demokratie (Diskursräume, Datenschutz, Rechenschaftspflicht/Verantwortung)

nach Maja Göpel, 2023, [https://media.ccc.de/v/37c3-12324-on\\_digitalisation\\_sustainability\\_climate\\_justice](https://media.ccc.de/v/37c3-12324-on_digitalisation_sustainability_climate_justice)

# Planetare Grenzen



© Steffen et al. 2015, übersetzt

- |   |   |
|---|---|
| ■ sicherer Handlungsräum verlassen; hohes Risiko gravierender Folgen    | ■ Menschheit agiert im sicheren Handlungsräum |
| ■ sicherer Handlungsräum verlassen; erhöhtes Risiko gravierender Folgen | ■ Belastbarkeitsgrenze nicht definiert        |



The image shows the front cover of a report. At the top right, the acronym "WBGU" is written in large, bold, blue letters. Below it, in smaller white text, is the full name: "Wissenschaftlicher Beirat der Bundesregierung Globale Umweltveränderungen". To the left of the acronym, the word "Hauptgutachten" is printed in large, white, sans-serif capital letters. In the center, the title "Unsere gemeinsame digitale Zukunft" is displayed in a large, white, bold, sans-serif font. The bottom half of the cover features a photograph of a diverse group of children wearing virtual reality headsets, looking upwards and to the right. They are outdoors, with green trees visible in the background. The overall color scheme is a bright blue for the top half and a natural green and white for the bottom half.

**ipcc**  
INTERGOVERNMENTAL PANEL ON **climate change**

<https://publication2023.bits-und-baeume.org/>

<https://www.wbgu.de/de/publikationen/hauptgutachten>

<https://www.de-ipcc.de/>

# What's next?

|    |           |  |   |
|----|-----------|--|---|
| 13 | 29.01.24  | Technik-, Informations- und informatische Berufsethik II, Informatik und Nachhaltigkeit [EG/FP]<br><b>(in Präsenz)</b> | <ol style="list-style-type: none"> <li>1. Existierende „Codes“ der informatischen „Berufsethik“ und Rolle der „Technikfolgenabschätzung“</li> <li>2. Ausgewählte weitere Themen (z.B. Value-based Engineering)</li> <li>3. Aktuelle Diskussionen und Stellungnahmen (z.B. Ethikrat, ...)</li> <li>4. tbd</li> </ol> |
|    | 01.02.24  | Poster-Feedback<br><b>(in Präsenz)</b>   | Im üblichen Hörsaal – für Gruppen aus Poster-Woche 1  |
| 14 | 05.02.24  | Poster-Feedback<br><b>(in Präsenz)</b>   | Im üblichen Hörsaal – für Gruppen aus Poster-Woche 2  |
|    | 08.02.24  | Freie Themenwahl & Roundup [FP]<br><b>(ausnahmsweise in Präsenz)</b>   | <p>Abschlussveranstaltung:</p> <ul style="list-style-type: none"> <li>• Wunschthemen</li> <li>• ISE beyond IG</li> <li>• ...</li> </ul>   |
| 15 | 12./15.02 | Ggf. Open Consultation Semester-Essay (tbd)<br><b>(via Zoom)</b>   | <b>Abgabe Essay: Donnerstag, 22.02.24, 23:59</b>  |

**Let us know!**

fin