UNIVERSITÄT DES SAARLANDES

SIC Saarland Informatics Campus

# Transparenz und Erklärbarkeit algorithmischer Entscheidungssysteme

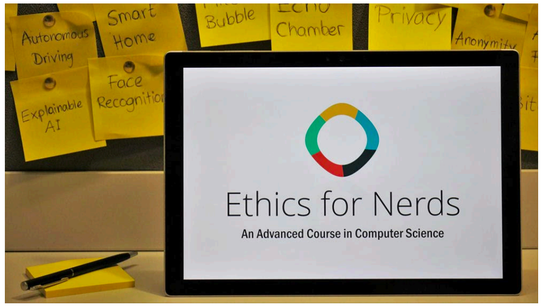
## Begründung, Sorten, Herausforderungen

Explainable Intelligent Systems

ALGO RIGHT

**37. Konferenz der Informationsfreiheitsbeauftragten (IFK)**

**12. Juni 2019**




Als Dozent: *Ethics for Nerds*
(Vertiefungsvorlesung)
seit 2016

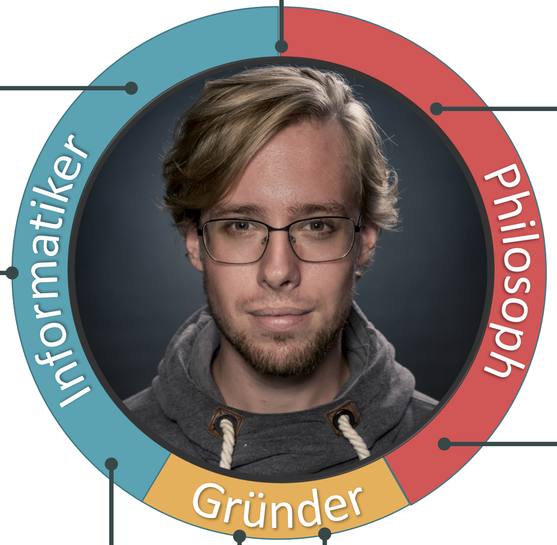
Forschung und
Dissertationsprojekt (seit 2018)
zu Maschinenerklärbarkeit

Explainable Intelligent Systems

VolkswagenStiftung

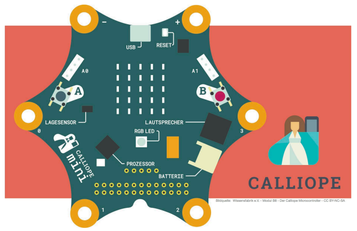
Bachelor (2011) und Master
(2013) in Informatik an der UdS

Seit 2015 Wissenschaftlicher
Mitarbeiter an der Dependable
Systems Group von
Professor Holger Hermanns




Master (2014) in
Philosophie an der UdS

Seit 2015
Wissenschaftlicher
Mitarbeiter am
Lehrstuhl für
Praktische
Philosophie von
Professorin Ulla
Wessels und
Professor Christoph
Fehige




Projekt mit Professorin Verena Wolf zur
Erforschung didaktischer Konzepte für
Digitale Bildung im Grundschulalter inkl.
Calliope

Mitgründer des
saarländischen
Think Tank für
gute Digitalisierung

ALGO RIGHT

eXIST
Existenzgründungen
aus der Wissenschaft

redkea

EXIST-Gründerstipendium
(2016 – 2017)

Als Dozent: Seminare
zu Normativer Ethik,
Angewandte Ethik,
vor allem
Computerethik

Forschung und
Dissertationsprojekt
zur Ethik kollektiven
Handelns

Digitale Verwaltung

# Das Amt und meine Daten

Dank Digitalisierung soll die Verwaltung in Österreich bequemer und effizienter werden. Doch schon jetzt zeigt sich: Mit dieser technischen Revolution in den Amtsstuben tauchen auch neue Probleme auf.

Von Anja Reiter, 11. Mai 2019, 10:44 Uhr / ZEIT Österreich Nr. 20/2019, 9. Mai 2019 / 77 Kommentare

https://www.zeit.de/2019/20/digitale-verwaltung-behoerden-aemter-effizienzsteigerung-probleme

NETZPOLITIK.ORG

Überwachung

# Big Data bei der Polizei: Hessen sucht mit Palantir-Software nach Gefährdern

Hessen läutet eine grundlegende Veränderung der Polizeiarbeit in Deutschland ein: Eine Software von Palantir verknüpft Datenbestände neu, wertet sie aus und soll etwa sogenannte Gefährder identifizieren. Dies ist nicht nur rechtlich fragwürdig, sondern dürfte weitreichende Folgen haben, schreibt der Kriminologe Tobias Singelnstein im Grundrechte-Report 2019.

03.06.2019 um 12:00 Uhr - Gastbeitrag, Tobias Singelnstein - eine Ergänzung

https://netzpolitik.org/2019/big-data-bei-der-polizei-hessen-sucht-mit-palantir-software-nach-gefaehrdern/

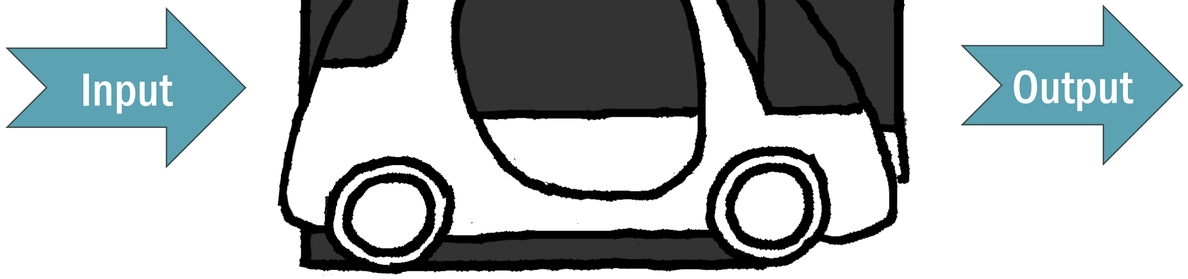
# Black Box: Was ist das?

## **Algorithmus**/Computerprogramm

```
                                (* Schritt3:                *)
WHILE A <> B DO                 (* Solange A ungleich B     *)
BEGIN
                                (* Schritt1:                *)
    IF B > A THEN               (* Falls B größer als A     *)
    BEGIN
        H := A; A := B; B := H  (* A und B vertauschen      *)
    END;
                                (* Schritt2:                *)
    A := A - B                  (* A durch A-B ersetzen     *)
END;
```

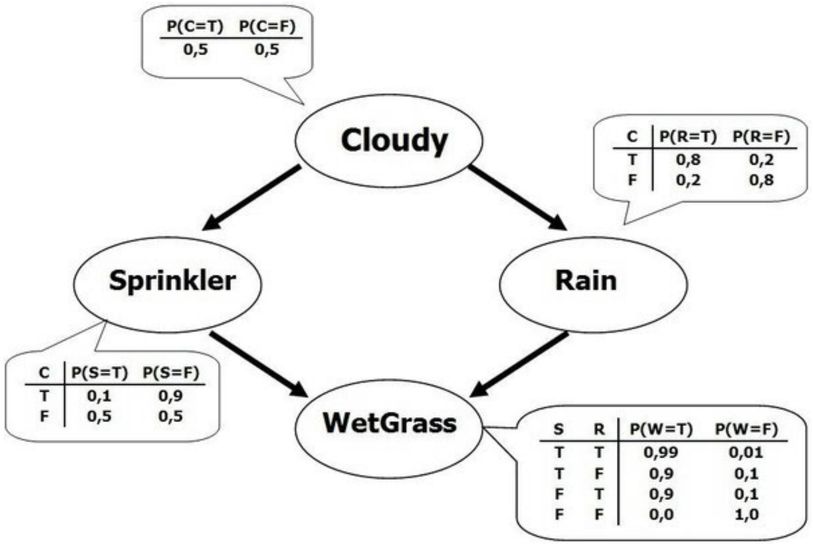
## Spezielle Sorte von ***Künstlicher Intelligenz*** (KI): *Lernende Software/Machine Learning (ML)*

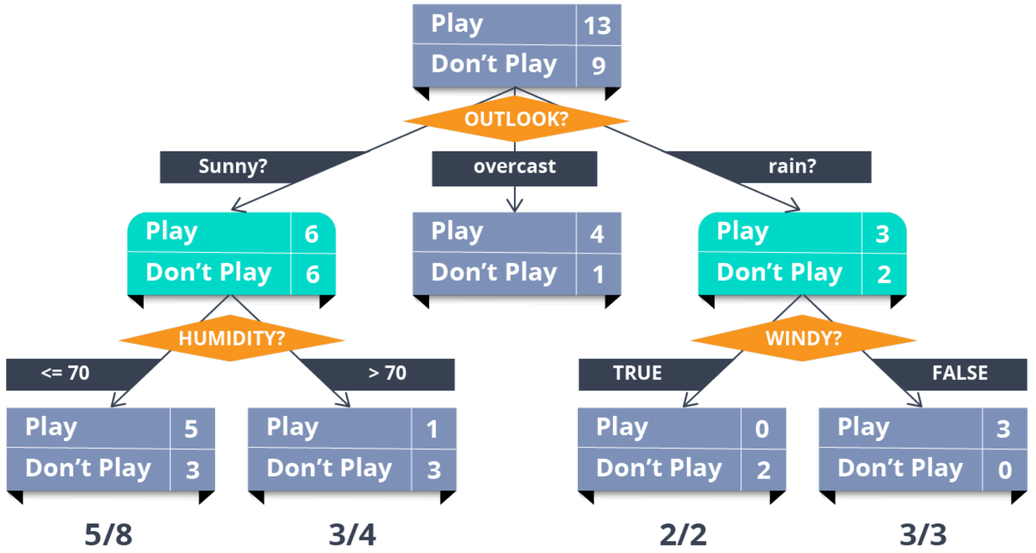
VS.

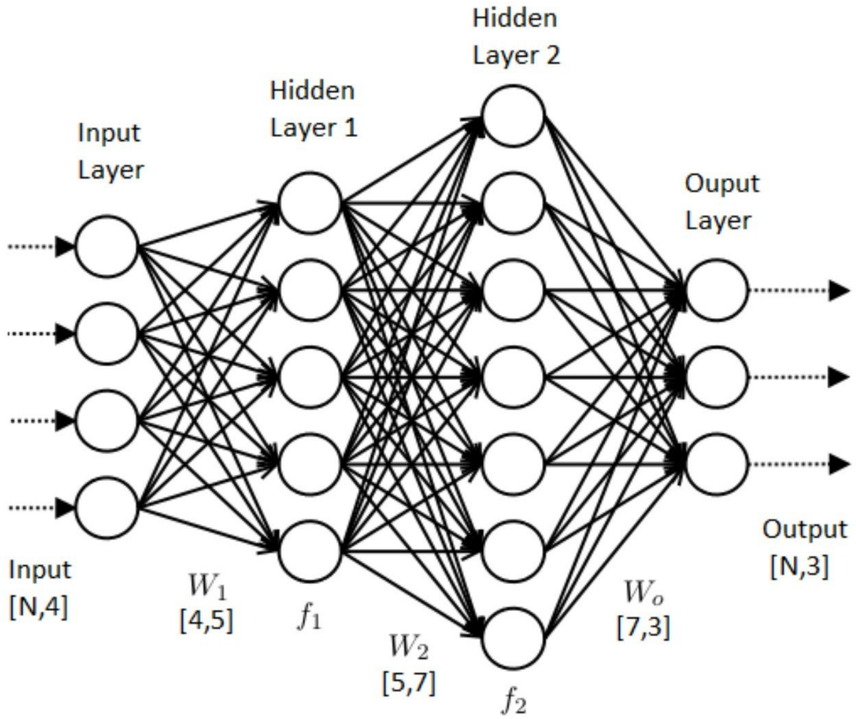



1. System ≠ Software ≠ Algorithmen ≠ KI ≠ Machine Learning
2. Das Stück Software nennt man gemeinhin „Modell“
3. Unterschiedliche Gründe für Modelle als Black Box: praktische & theoretische

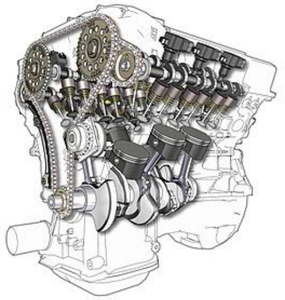
```
53
54      uppercase_sample = 'ABCDEFGHIJKLMNOPQRSTUVWXYZ'
55      lowercase_sample = 'abcdefghijklmnopqrstuvwxyz'
56      digit_sample = '0123456789'
57      if keras.backend.image_dim_ordering() != 'tf' and count <= 100:
58          keras.backend.set_image_dim_ordering('tf')
59          print("INFO: '~/.keras/keras.json' sets 'image_dim_ordering' to "
60                "'th', temporarily setting to 'tf'")
61
62      # Create TF session and set as Keras backend session
63      sess = tf.Session()
64      keras.backend.set_session(sess)
65
66      # Get MNIST test data
67      X_train, Y_train, X_test, Y_test = data_mnist(train_start=train_start,
68                                                    train_end=train_end,
69                                                    test_start=test_start,
70                                                    test_end=test_end)
71
72      assert Y_train.shape[1] == 10
```



# „Daten sind das neue ÖL“ – ML der neue Motor?

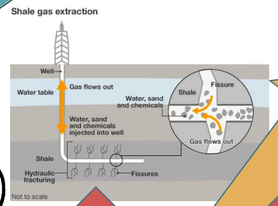


Immense Datenmengen (aka „Big Data“)

sind Basis für

Lern-algorithmen

produziert




KI

KI

Welt

Klassische Software

Klassische Software

KI

Resultiert in

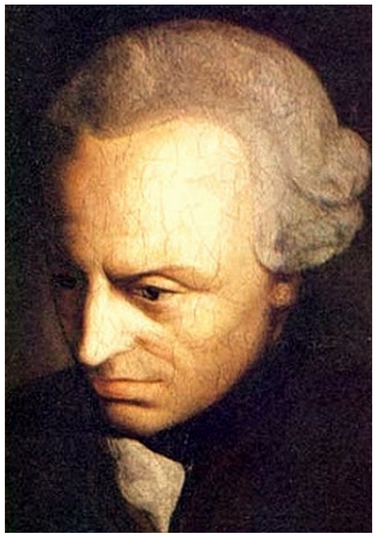
beeinflusst/ist wirksamer Teil von

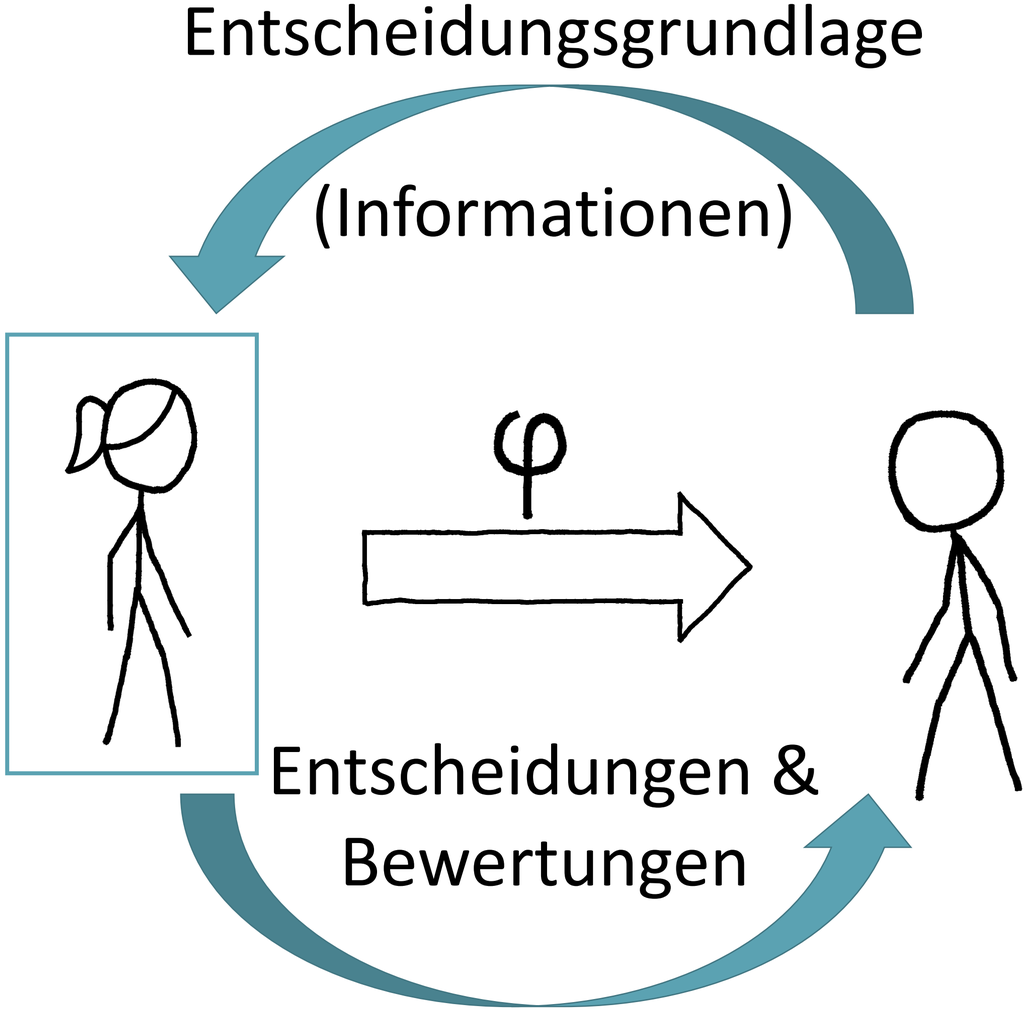
KI-System

# Subjekt und Objekt: Klassisch

Bislang:

- Menschen
  - erfahrungsbasiert vs. regelbasiert
  - alltäglich vs. institutionell



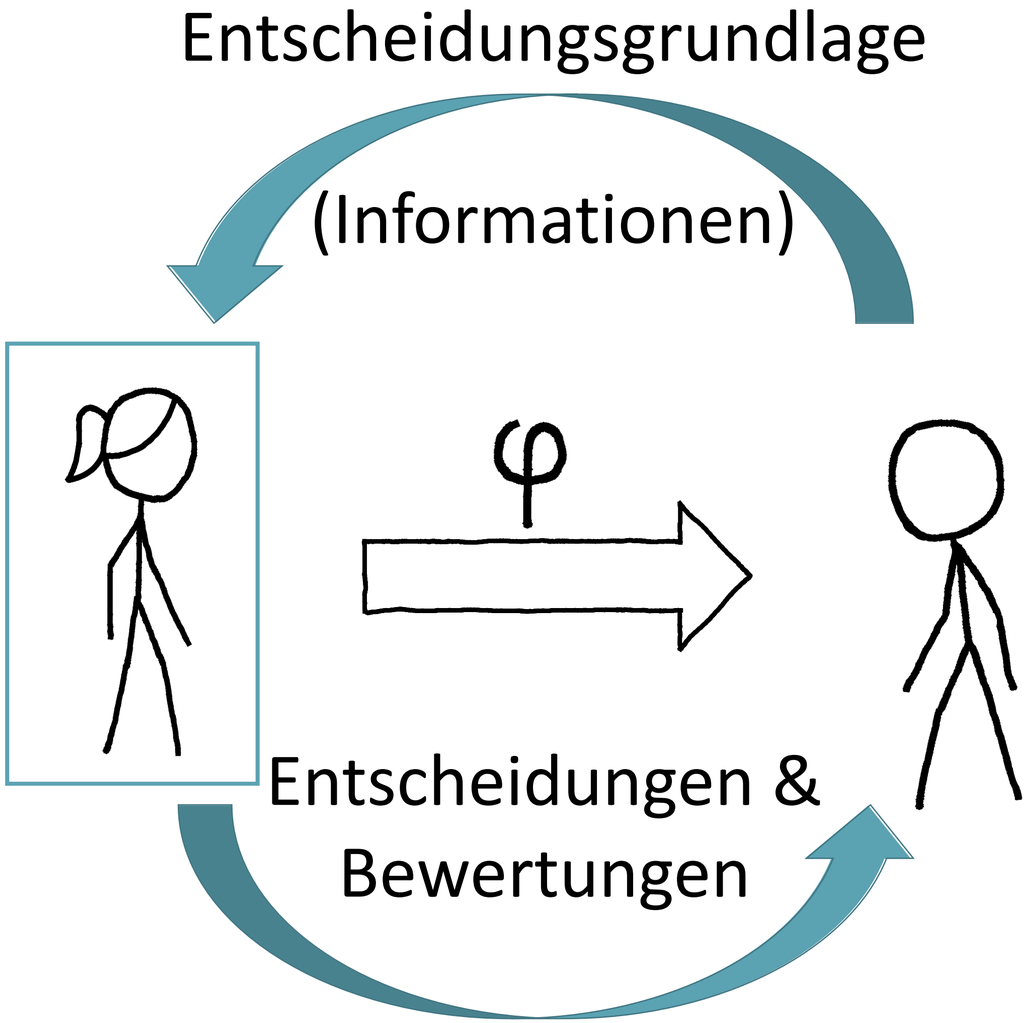


# Subjekt und Objekt: Klassisch

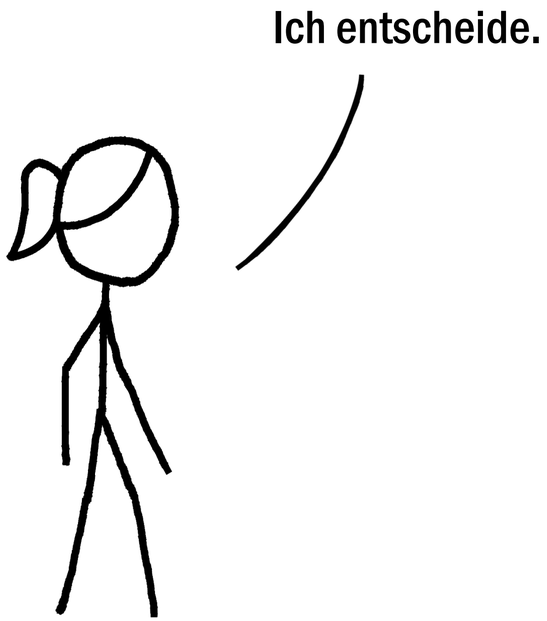
Bislang:

- Menschen
  - erfahrungsbasiert vs. regelbasiert
  - alltäglich vs. institutionell

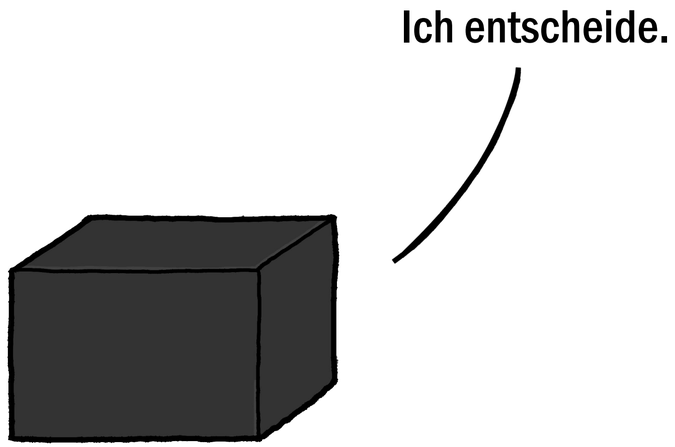




# Ein (unvermeidlicher?) Trend?







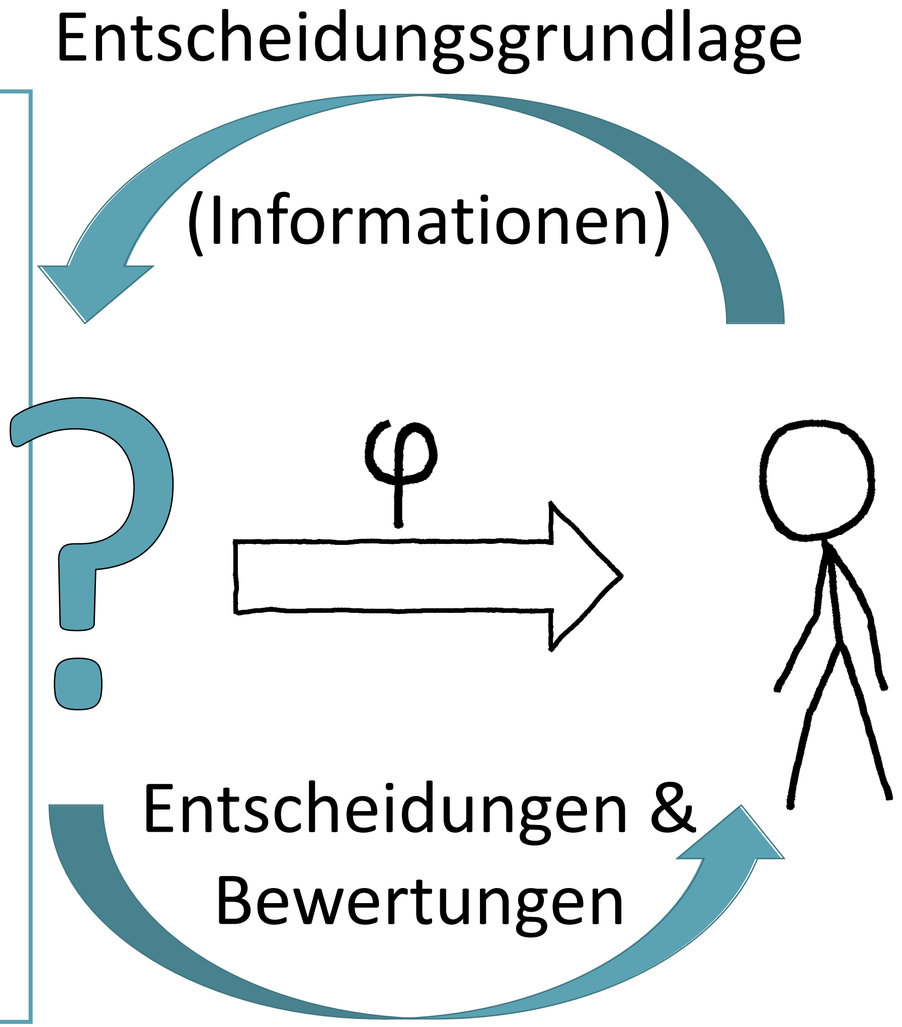
# Subjekt und Objekt

Bislang:

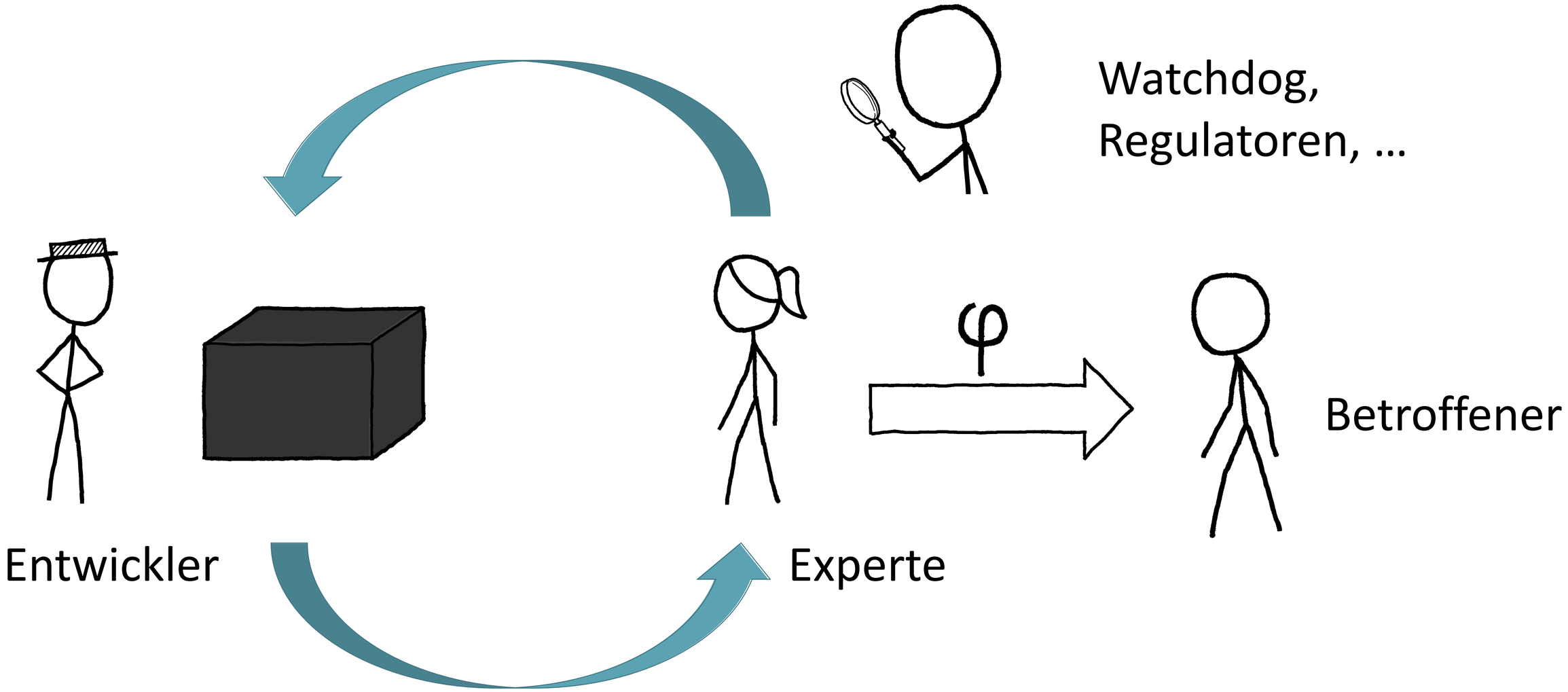
- Menschen
  - erfahrungsbasiert vs. regelbasiert
  - alltäglich vs. institutionell

Neuland:

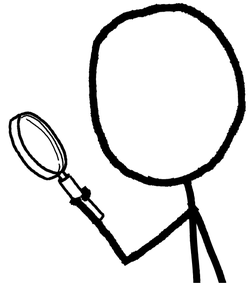
- Menschen, unterstützt von Algorithmen
- Algorithmen allein



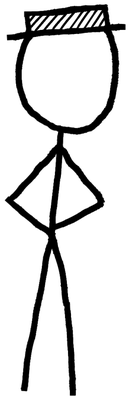
# Perspektiven

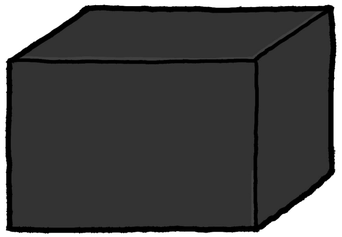


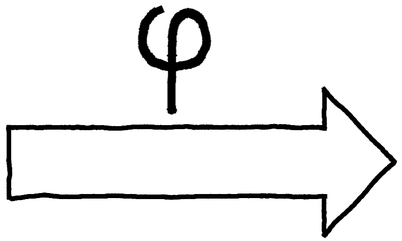
# Perspektiven



Watchdog,
Regulatoren, …



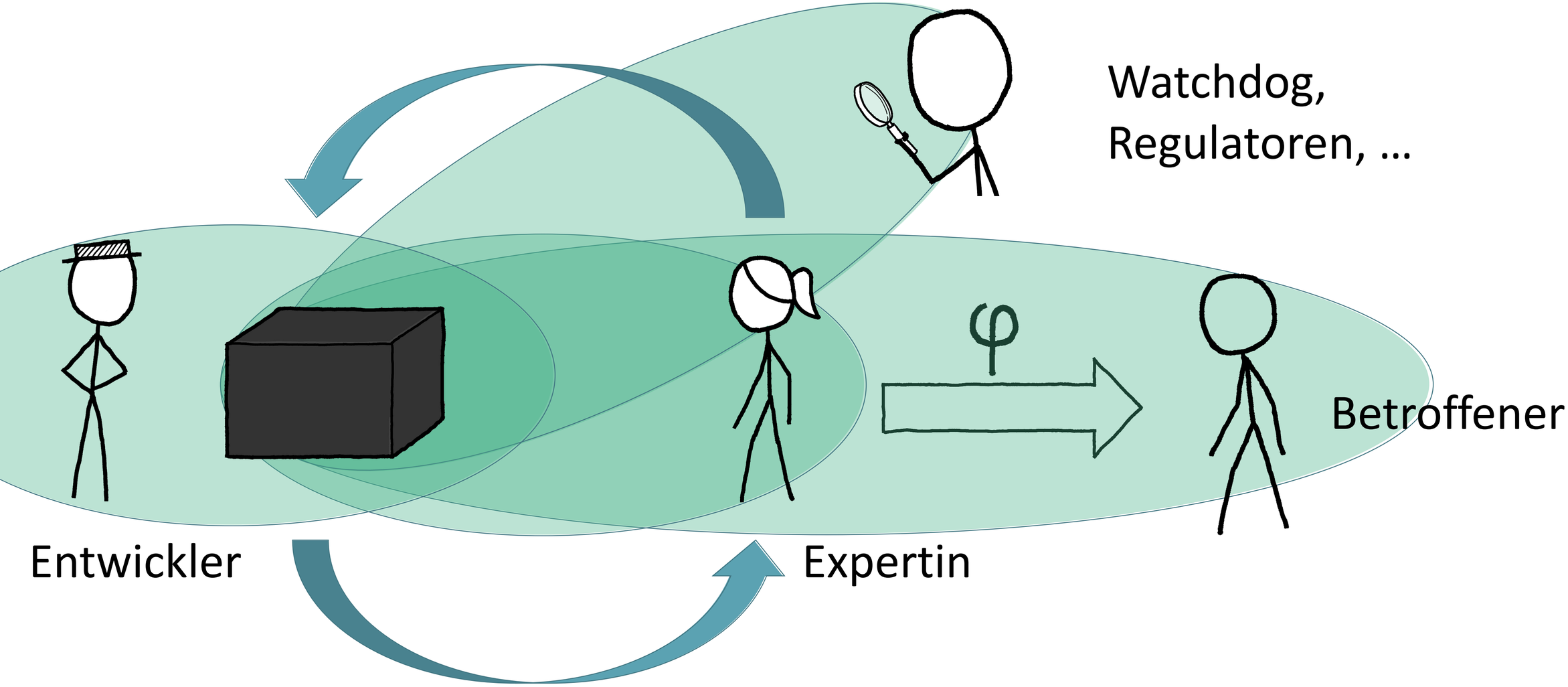





Betroffener

Entwickler

# Perspektiven



# Typ 1:
# Right to be Informed

## Suitable GDPR articles

**Art. 12 GDPR** Transparent information, communication and modalities for the exercise of the rights of the data subject, **Art. 13 GDPR** Information to be provided where personal data are collected from the data subject, **Art. 14 GDPR** Information to be provided where personal data have not been obtained from the data subject

https://gdpr-info.eu/issues/right-to-be-informed/

# Nicht: Right to Explanation

- Nicht verwechseln mit dem *Right to Explanation*.
  Das ist nur Willensbekundung, Teil eines Erwägungsgrundes.
- Wir wüssten auch gar nicht, wie das gehen soll, zumindest nicht bei ‚modernen' Verfahren. (Später mehr dazu.)
- Zum Teil aber womöglich impliziertes Recht.

## Right to Explanation: a Right that Never Was (in GDPR)

*Posted on 1st March 2018 by Eve the Analyst*

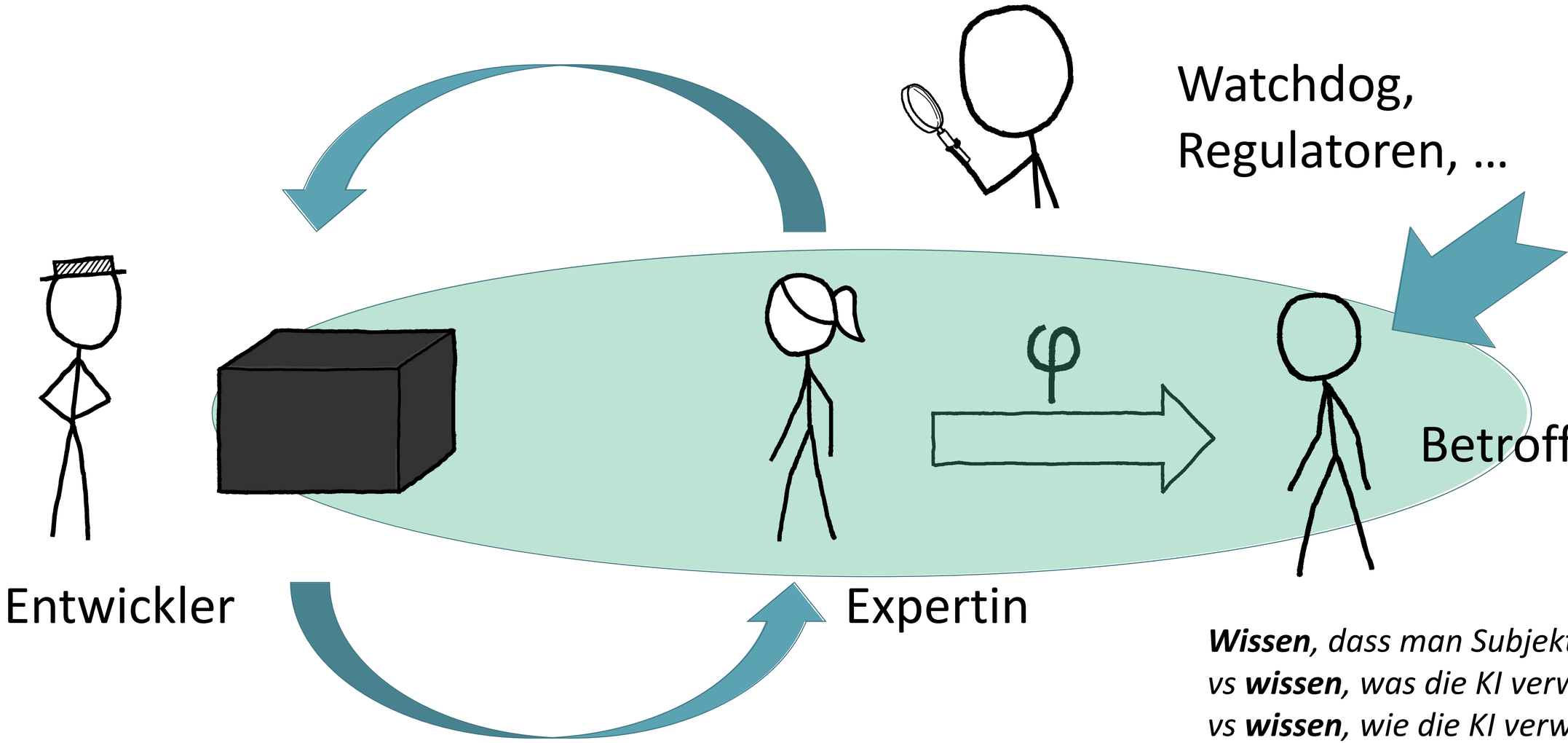
http://datawanderings.com/2018/03/01/right-to-explanation/

**Recital 71**
**EU GDPR**

*In any case, such processing should be subject to suitable safeguards, which should include* ***specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision*** *reached after such assessment and* ***to challenge the decision.***

https://www.privacy-regulation.eu/en/r71.htm

# Typ-1-Transparenz: Relevante Perspektiven




***Wissen**, dass man Subjekt ist*
*vs **wissen**, was die KI verwendet*
*vs **wissen**, wie die KI verwendet wird*
*vs **wissen**, wie die KI funktioniert*

# Typ-1-Transparenz

## Typ 1:
## Right to be Informed

*Umsetzbarkeit?* Höchstens eine regulatorische Herausforderung

*Warum wichtig?* Ohne Kenntnis, kein Hinterfragen bzw. Einspruch

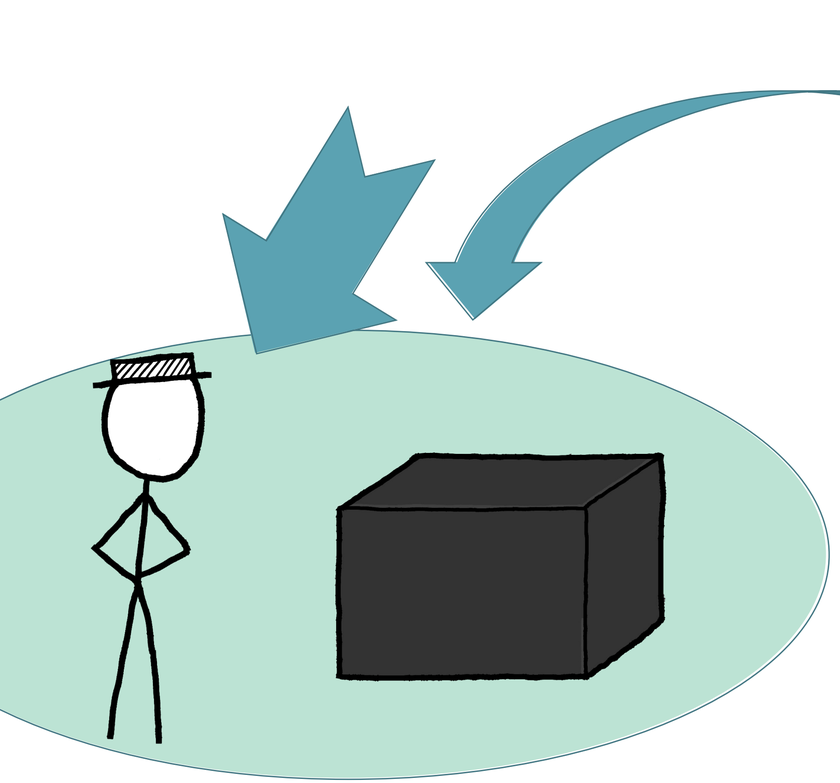
# Typ-2-Transparenz

Typ 2:

Transparenz durch
*grundlegendes Verstehen,*
*Testbarmachung* (*Validierbarkeit*) und
*formale Explikation*

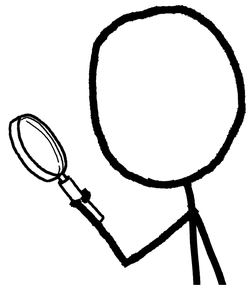
(auf Ebene des Systems und auf Ebene seiner Entscheidungen)

# Typ-2-Transparenz: Relevante Perspektiven



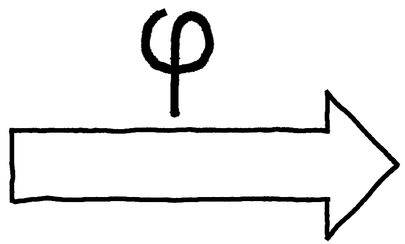
Entwickler

***Was baue ich da?***
***Tut es, was es soll?***
***Kann ich es testen?***
***Kann ich Teile davon wiederverwenden?***



Watchdog,
Regulatoren, …

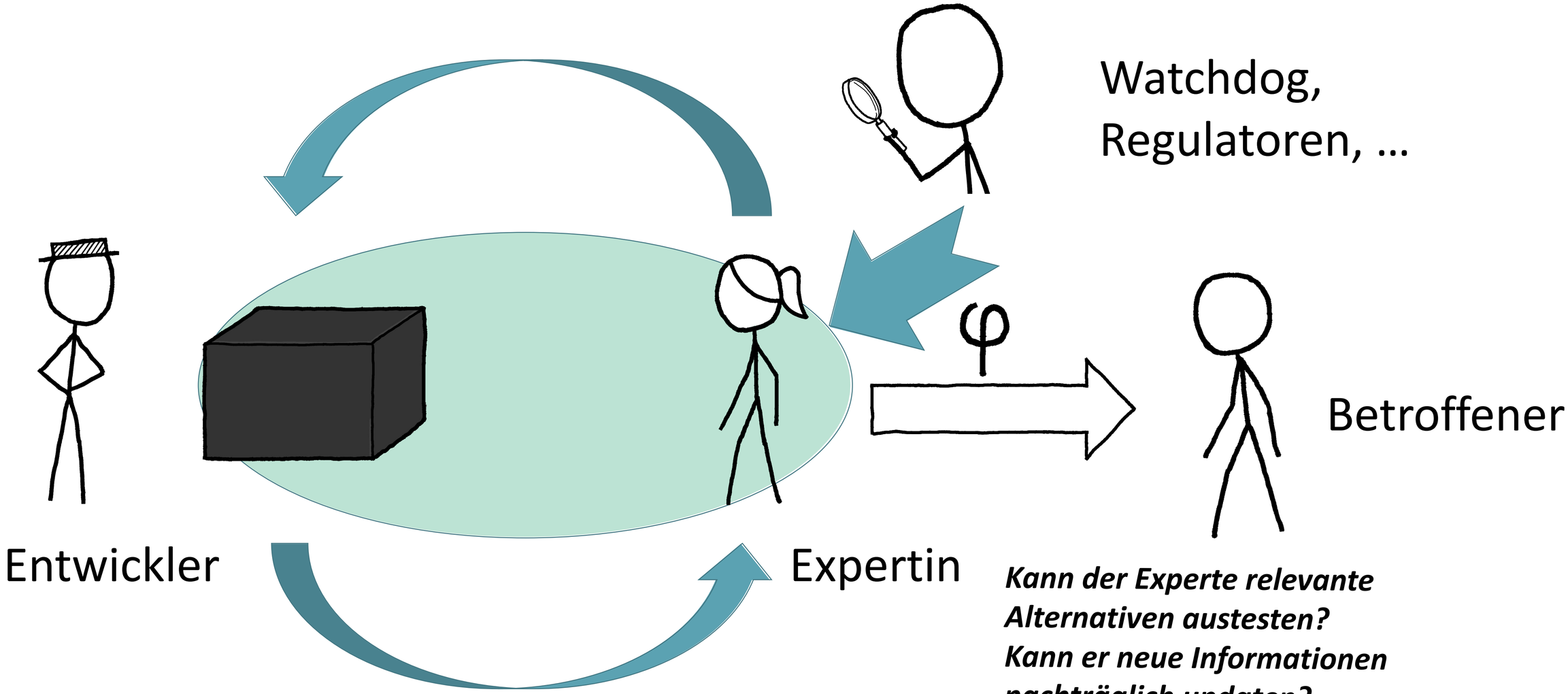
Expertin



Betroffener

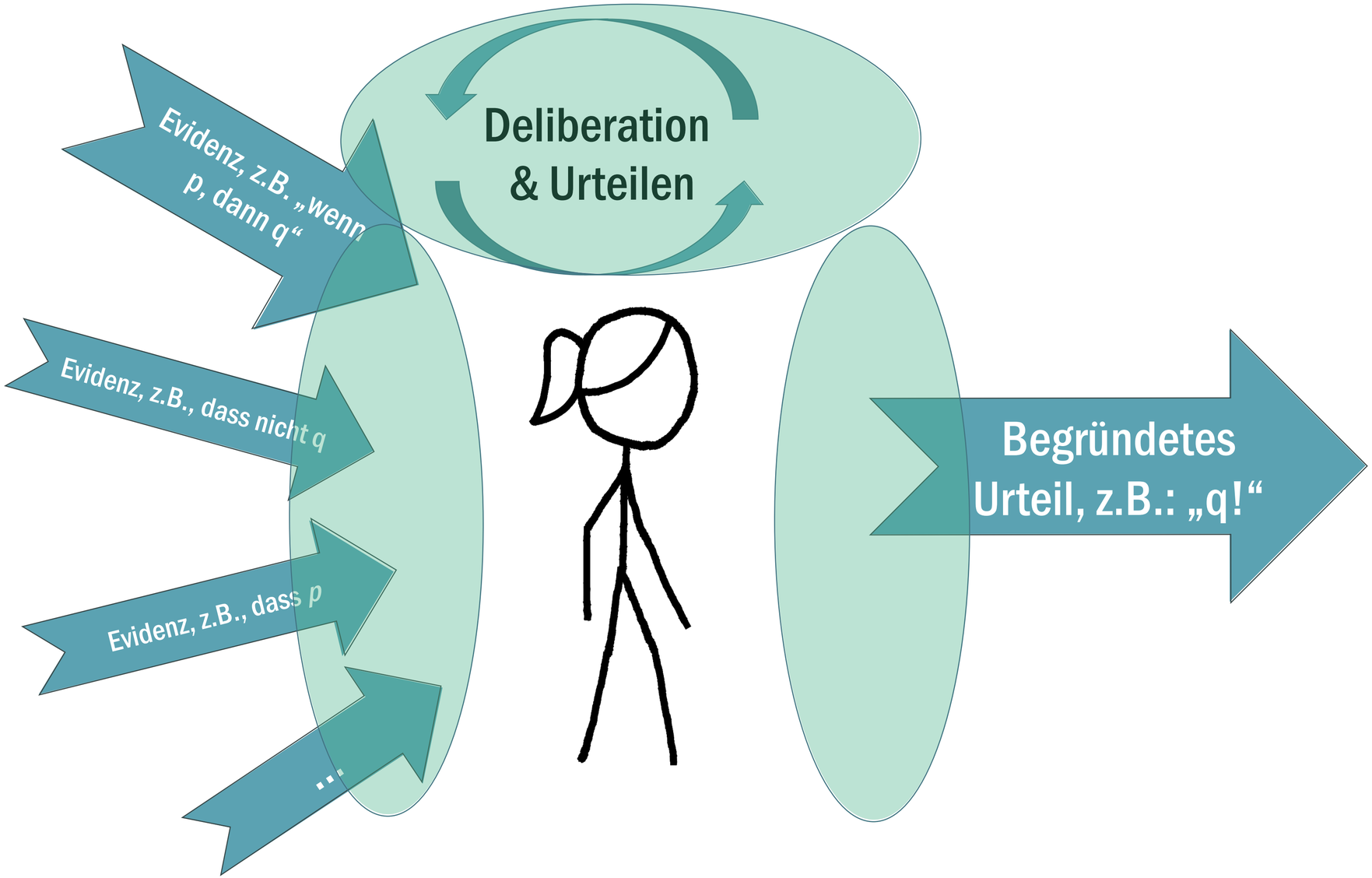
***(Wichtig, nicht-trivial für Machine-Learning.***
***Aber: keine Aufgabe der öffentlichen Hand)***

# Typ-3-Transparenz: Relevante Perspektiven

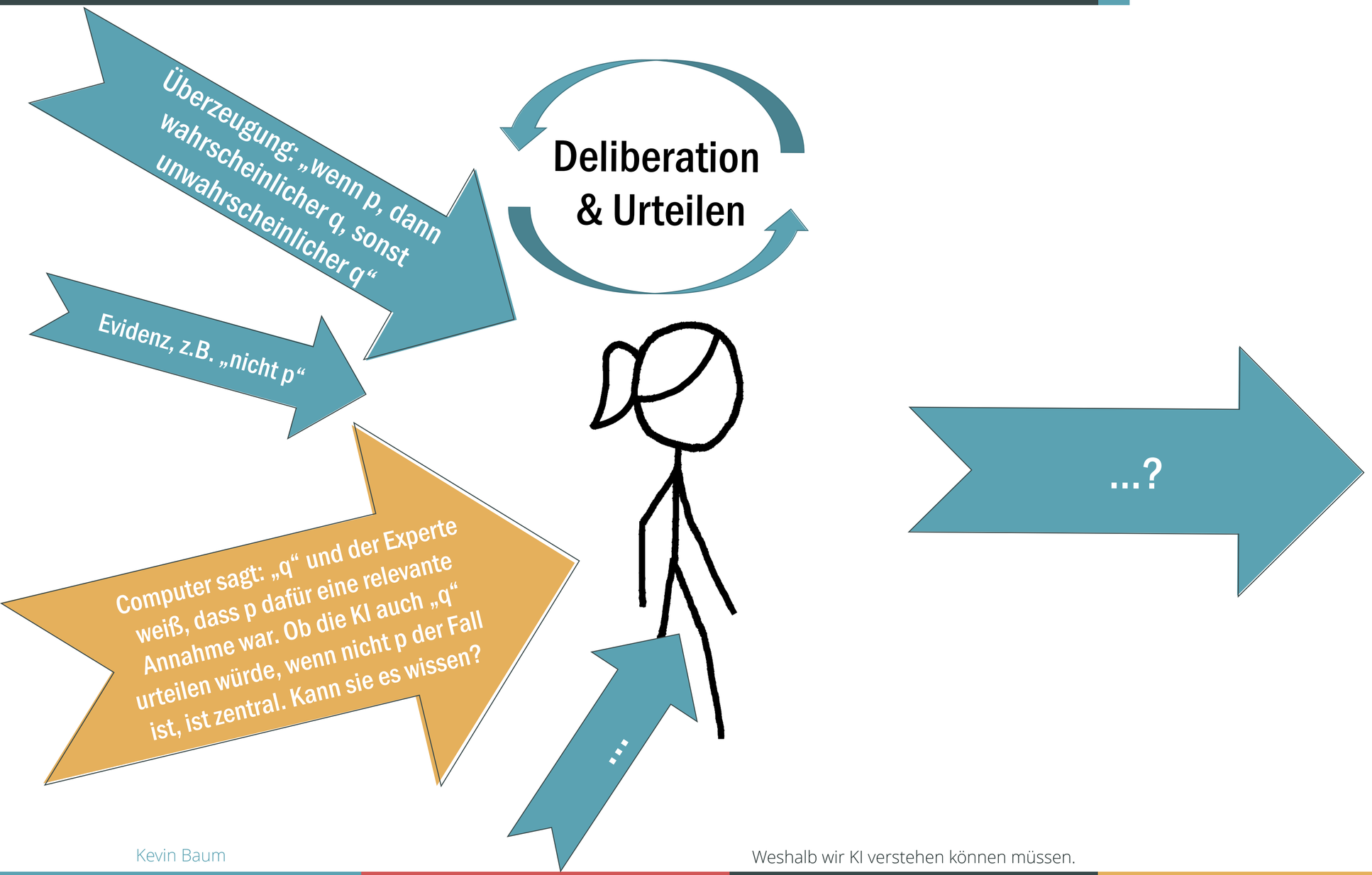



***Kann der Experte relevante Alternativen austesten?***
***Kann er neue Informationen nachträglich updaten?***
***Die Relevanz von Inputs erkennen?***

# Die Funktion menschlicher Entscheider



# Das ganz-oder-gar-nicht-Problem

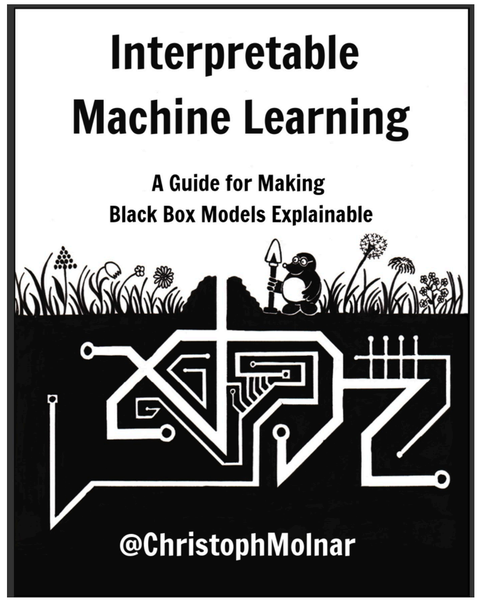



Abwägen und angemessen berücksichtigen, das kann die Expertin möglicherweise nicht.

Verbleibende Entscheidung wäre: „Folge dem Computer oder verwerfe dessen Output!“

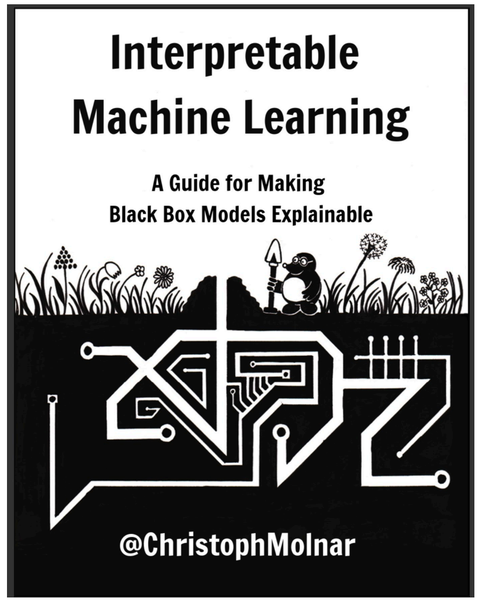
# Bedarf: Erklärungen

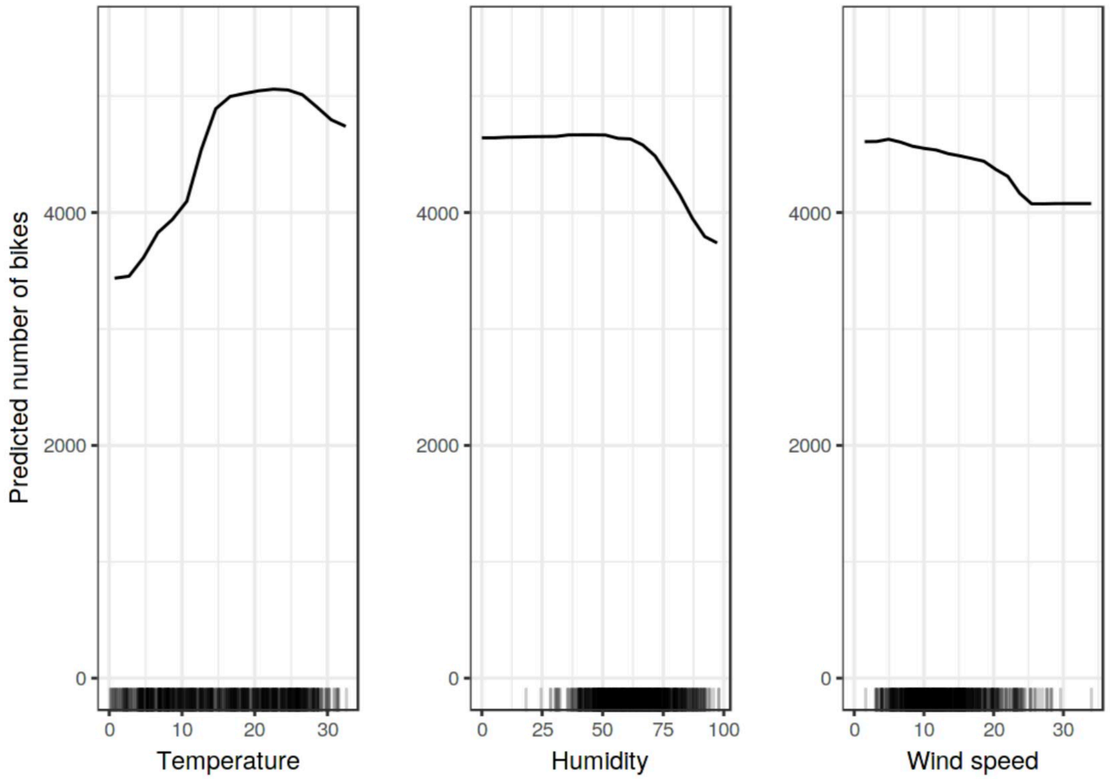
- Hier stellen sich notwendigerweise „Warum-Fragen“
  - „Warum bewertet es diesen Bewerber so gut/so schlecht?“
  - „Warum ist Bewerber A, der zu Minderheit X gehört, vor Bewerber B gerankt?“
  - …
- Warum-Fragen wollen als Antwort Erklärungen.
- Wir benötigen Antworten auf Erklärungen.

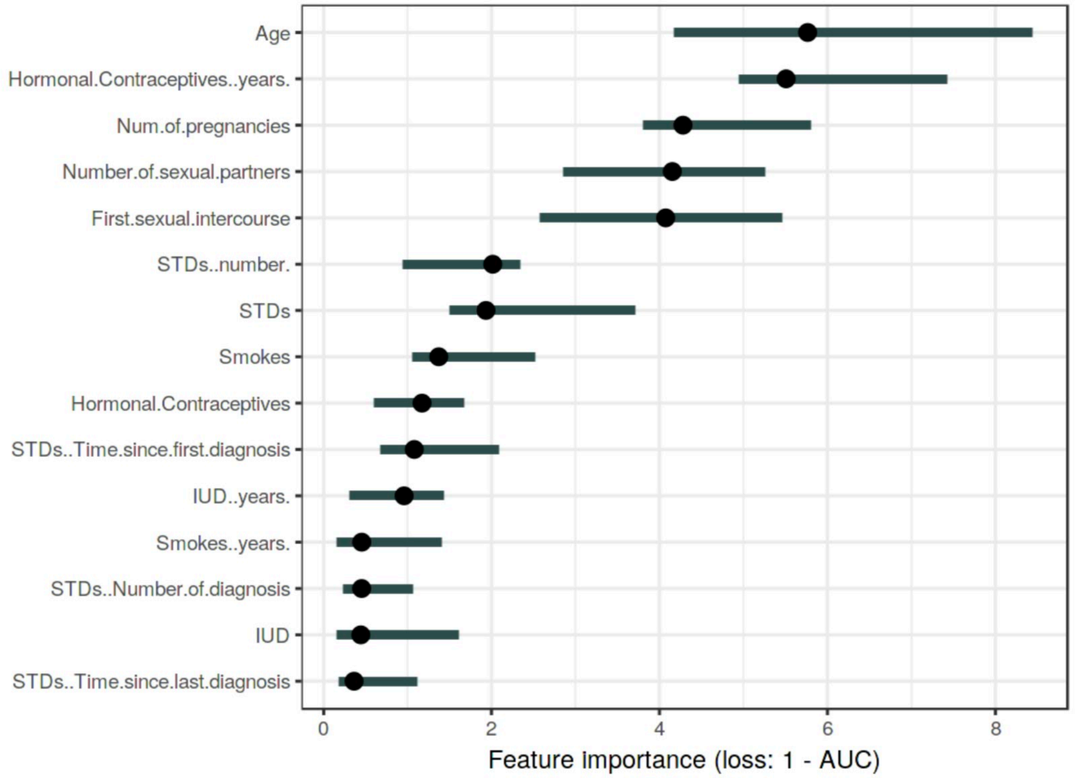


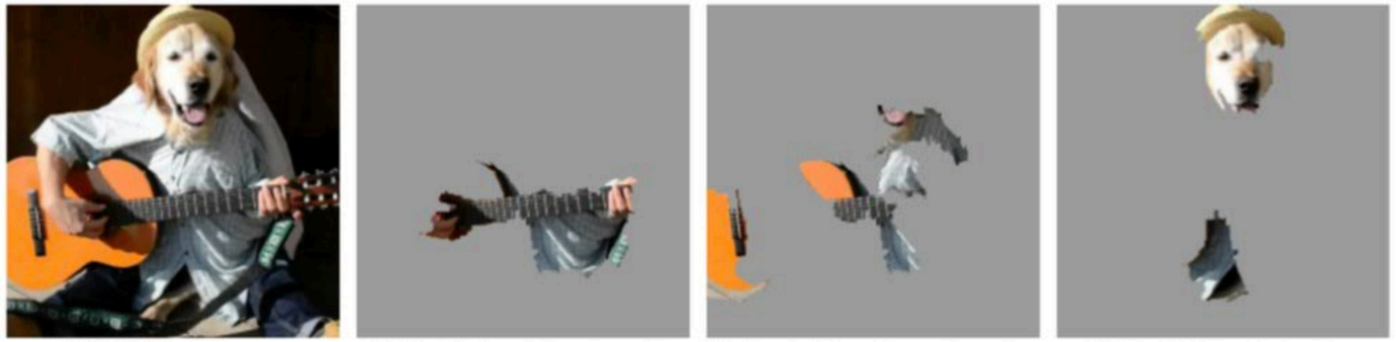
# Bedarf: Erklärungen

- Wir benötigen Antworten auf Erklärungen:
  Modell-agnostische Methoden vs. Transparenz by Design (erklärbare Modelle)
- Methoden: Nicht alle Antworten müssen dabei verbal sein.
  - Partial Dependence Plot (PDP)
  - Feature Interaction, Feature Importance
  - Global & Local Surrogates (LIME)
  - Example-Based:
    - Counterfactuals, Adversarial Examples, Prototypes & Crticisims
- Rein formale Erklärungen („Mathe“) nennt man *Explikationen*





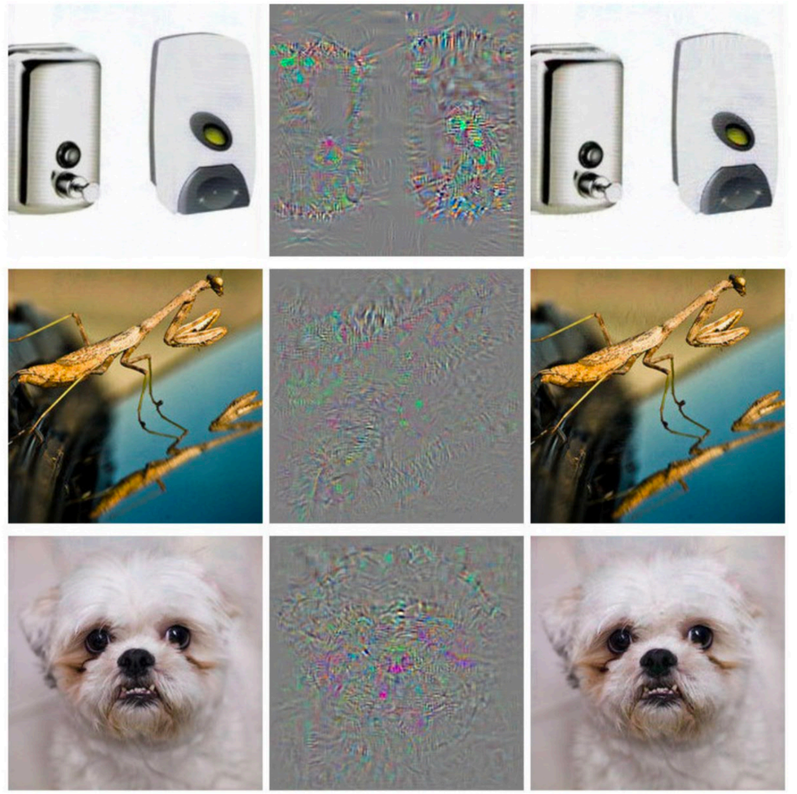



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

**LIME explanations for the top 3 classes for image classification made by Google's Inception neural network. The example is taken from the LIME paper (Ribeiro et. al., 2016).**


| Score | GPA | LSAT | Race | GPA x’ | LSAT x’ | Race x’ |
|---|---|---|---|---|---|---|
| 0.17 | 3.1 | 39.0 | 0 | 3.1 | 34.0 | 0 |
| 0.54 | 3.7 | 48.0 | 0 | 3.7 | 32.4 | 0 |
| -0.77 | 3.3 | 28.0 | 1 | 3.3 | 33.5 | 0 |
| -0.83 | 2.4 | 28.5 | 1 | 2.4 | 35.8 | 0 |
| -0.57 | 2.7 | 18.3 | 0 | 2.7 | 34.9 | 0 |



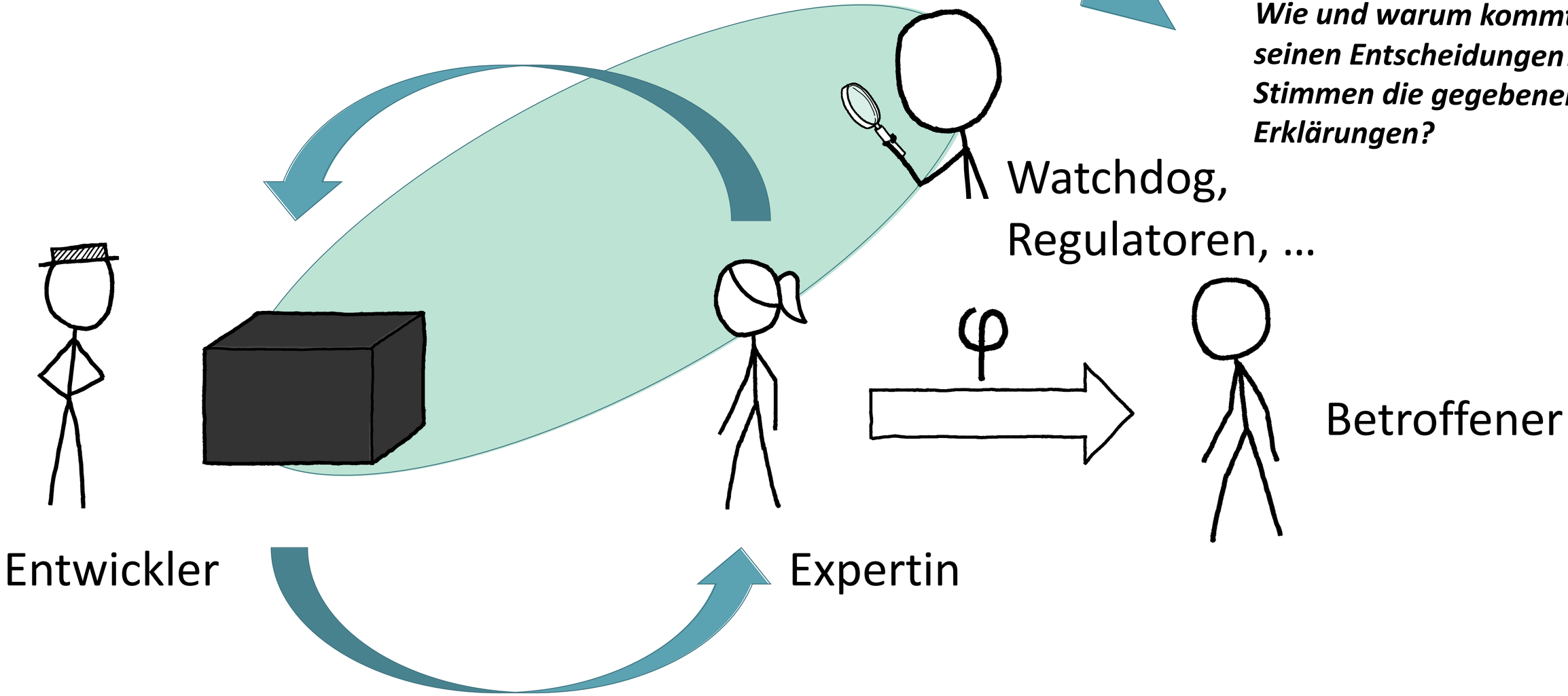
Adversarial examples for AlexNet by Szegedy et. al (2013). All images in the left column are correctly classified. The middle column shows the (magnified) error added to the images to produce the images in the right column all categorized (incorrectly) as ‘Ostrich’.

# Beschränkungen

- Nur manche Methoden funktionieren für Einzelfälle, d.h.
  - Nur manche Methoden erklären einen bestimmten Ausgabe gegeben eine bestimmte Eingabe. Beispielsweise *Counterfactuals*, *Feature Importance*
  - Viele Methoden geben eher *Einsicht in das Modell*: was es gut kann, was schlecht. Beispielsweise: *Adversarial Examples*, *Prototyps & Criticisms, Feature Interaction*
- Die Methoden bieten vor allem Einsichten für visuelle Inputs, Bilder & Videos.
  Nicht ideal für Verwaltungsaufgaben.
- Sonst braucht es oft eher mathematisches Know-How.
  Hilfreich für Informatiker und Menschen mit Mathe-Background,
  eher ungeeignet für Fachexperten (Sachbearbeiter, Richter)

# Typ-2-Transparenz: Relevante Perspektiven

***Erfüllt das System bestimmte Anforderungen? Z.B.: Benachteiligt es nicht nach Gruppenzugehörigkeit? (Fairness) Wie und warum kommt es zu seinen Entscheidungen? Stimmen die gegebenen Erklärungen?***



# Typ-2-Transparenz

Typ 2:

Transparenz durch *Grundlegendes Verstehen,*
*Testbarmachung* (*Validierbarkeit*) und *formale Explikation*

*Umsetzbarkeit?*

- Praktische Hürden (*Betriebs- und Geschäftsgeheimnisse, Aufwand*):
  Bedarf einer *vertrauenswürdigen*, kompetenten und unabhängigen Prüfstelle („Algorithmen-TÜV“), Regulierung
- Theoretisch u.U. sehr anspruchsvoll bis unmöglich
  (für moderne *Machine-Learning-Verfahren*)

# Typ-2-Transparenz

Typ 2:

Transparenz durch *Grundlegendes Verstehen,*
*Testbarmachung* (*Validierbarkeit*) und *formale Explikation*

*Warum wichtig?*

- *Scheinobjektivierung* (*Mathwashing*) nicht ‚auf den Leim' gehen
- Kontrolle von Modellen, z.B. Erkennen systematischer & statistischer Fehler
- Teilweise verbesserte Einschätzbarkeit von individuellen Entscheidungen, ermöglicht teilweise Evaluation und Verbesserung (Feedback)

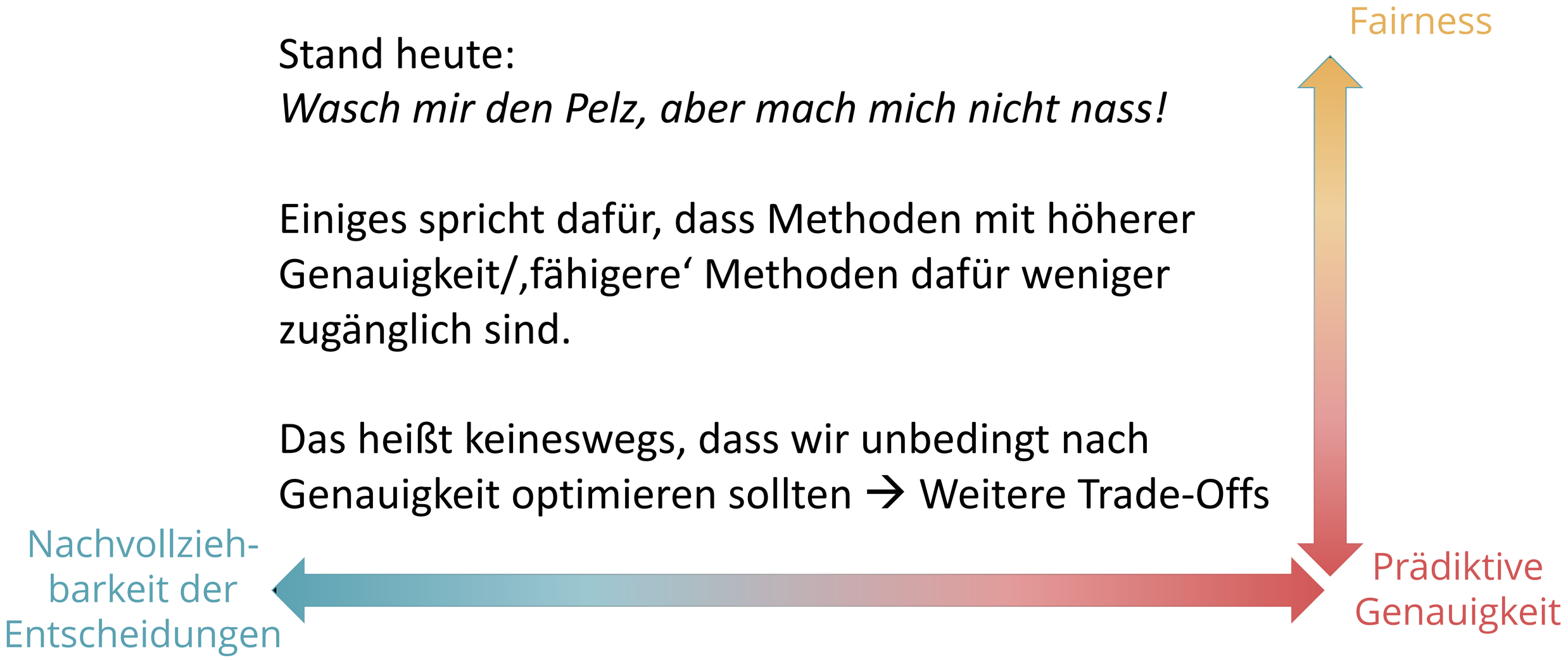
Solange nicht gegeben, dürfen solche Systeme *eigentlich* nicht verwendet werden.

# Trade-Offs

Stand heute:
*Wasch mir den Pelz, aber mach mich nicht nass!*

Einiges spricht dafür, dass Methoden mit höherer Genauigkeit/‚fähigere' Methoden dafür weniger zugänglich sind.

Das heißt keineswegs, dass wir unbedingt nach Genauigkeit optimieren sollten → Weitere Trade-Offs



# Typ-3-Transparenz

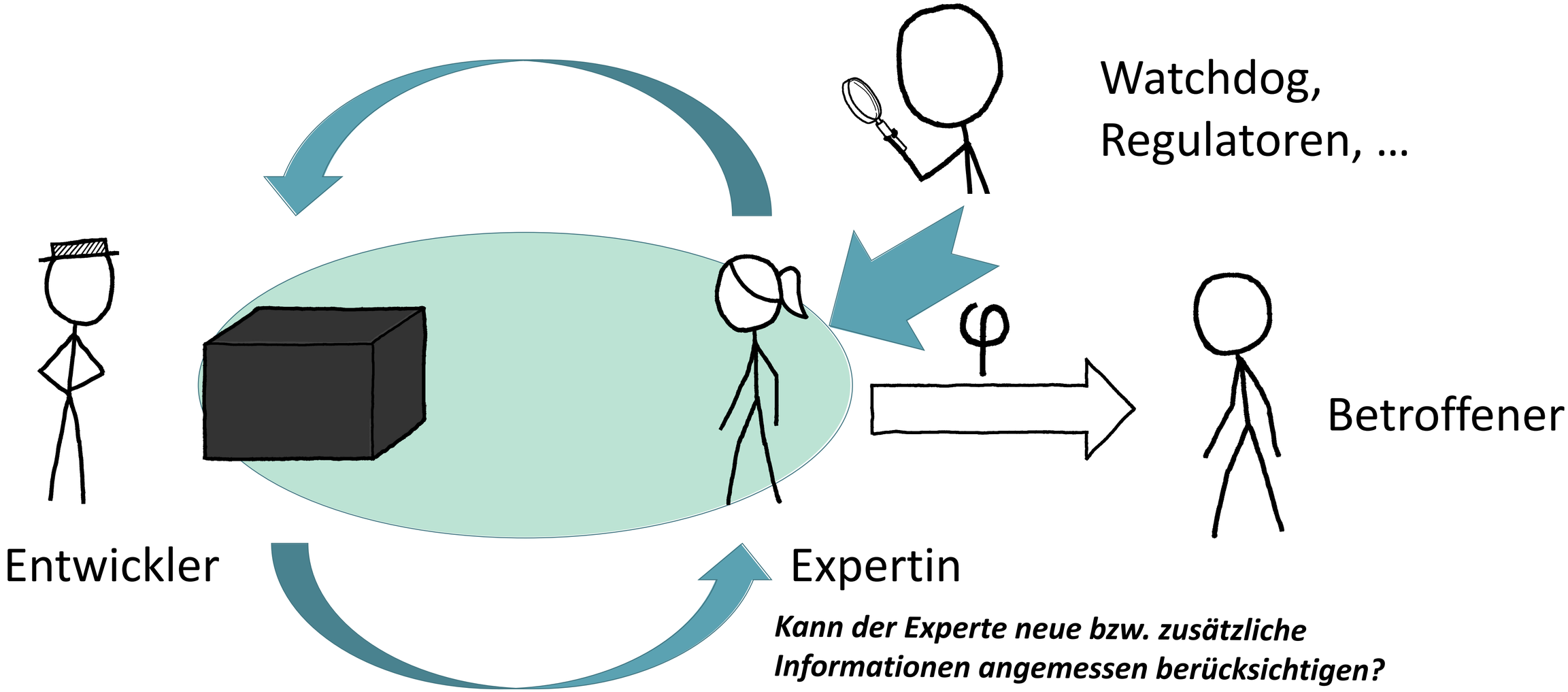

Typ 3:

Transparenz des Entscheidungsprozesses &
Zugang/Zugriff auf den internen, deliberativen Prozess

(Gründe, ihre Beziehung zueinander sowie ihre Gewichtung)

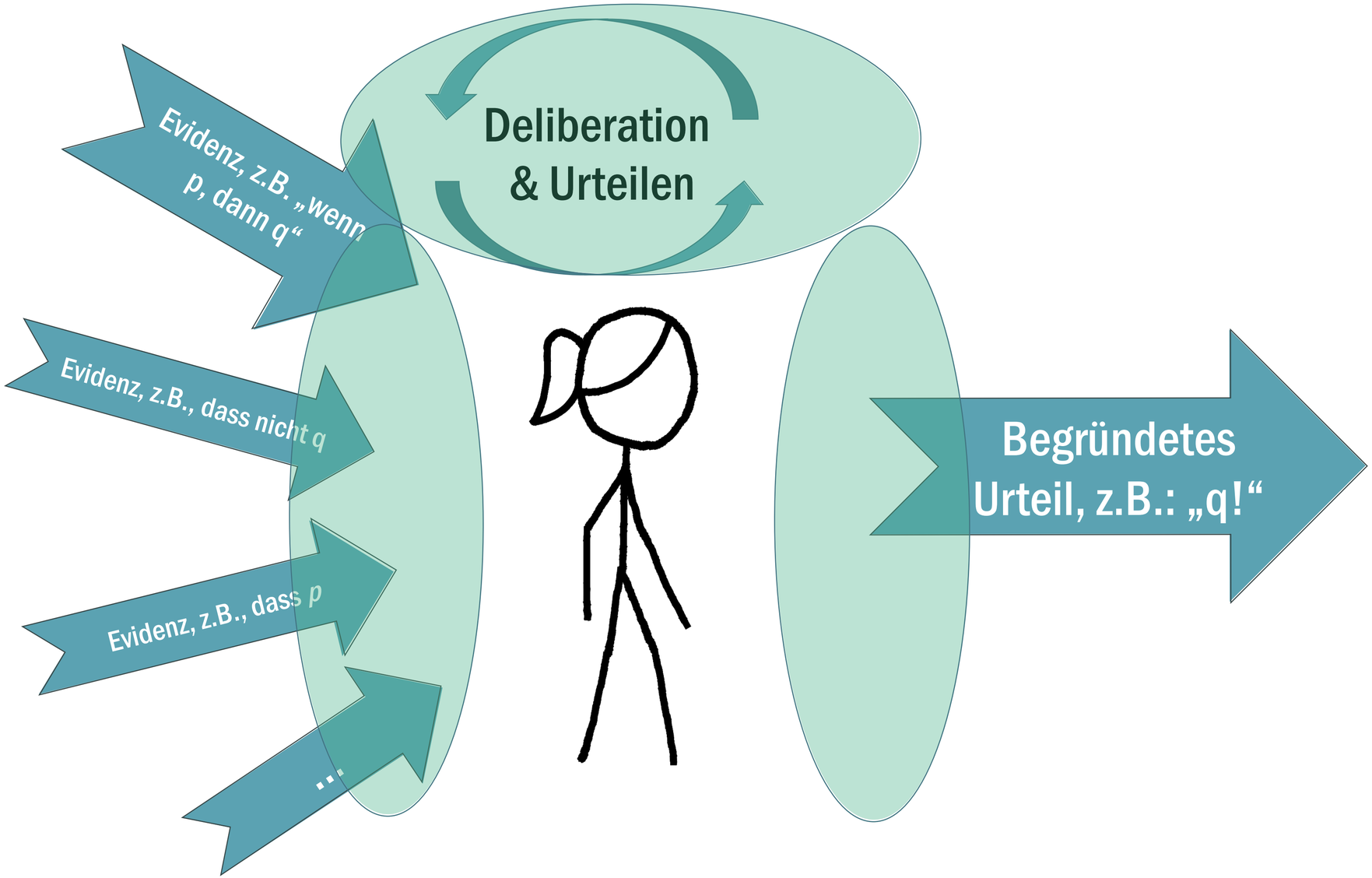
***Ziel: Individuelle Entscheidung verstehen und hinterfragbar machen***

# Typ-3-Transparenz: Relevante Perspektiven

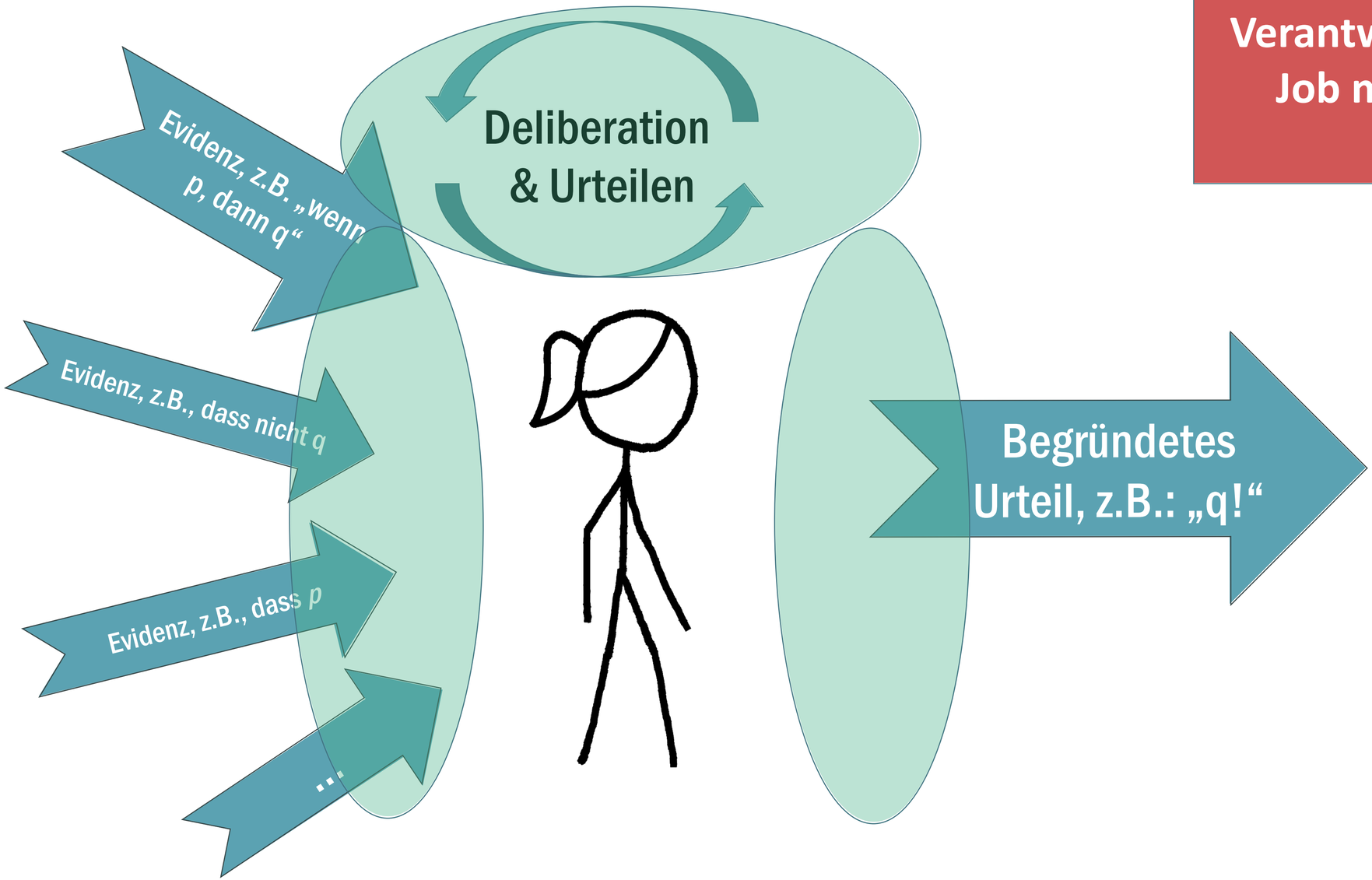



***Kann der Experte neue bzw. zusätzliche Informationen angemessen berücksichtigen? Kann er fehlerhafte Beurteilungen und Empfehlungen erkennen? (Filterfunktion!)***

# Die Funktion menschlicher Entscheider



# Die Funktion menschlicher Entscheider

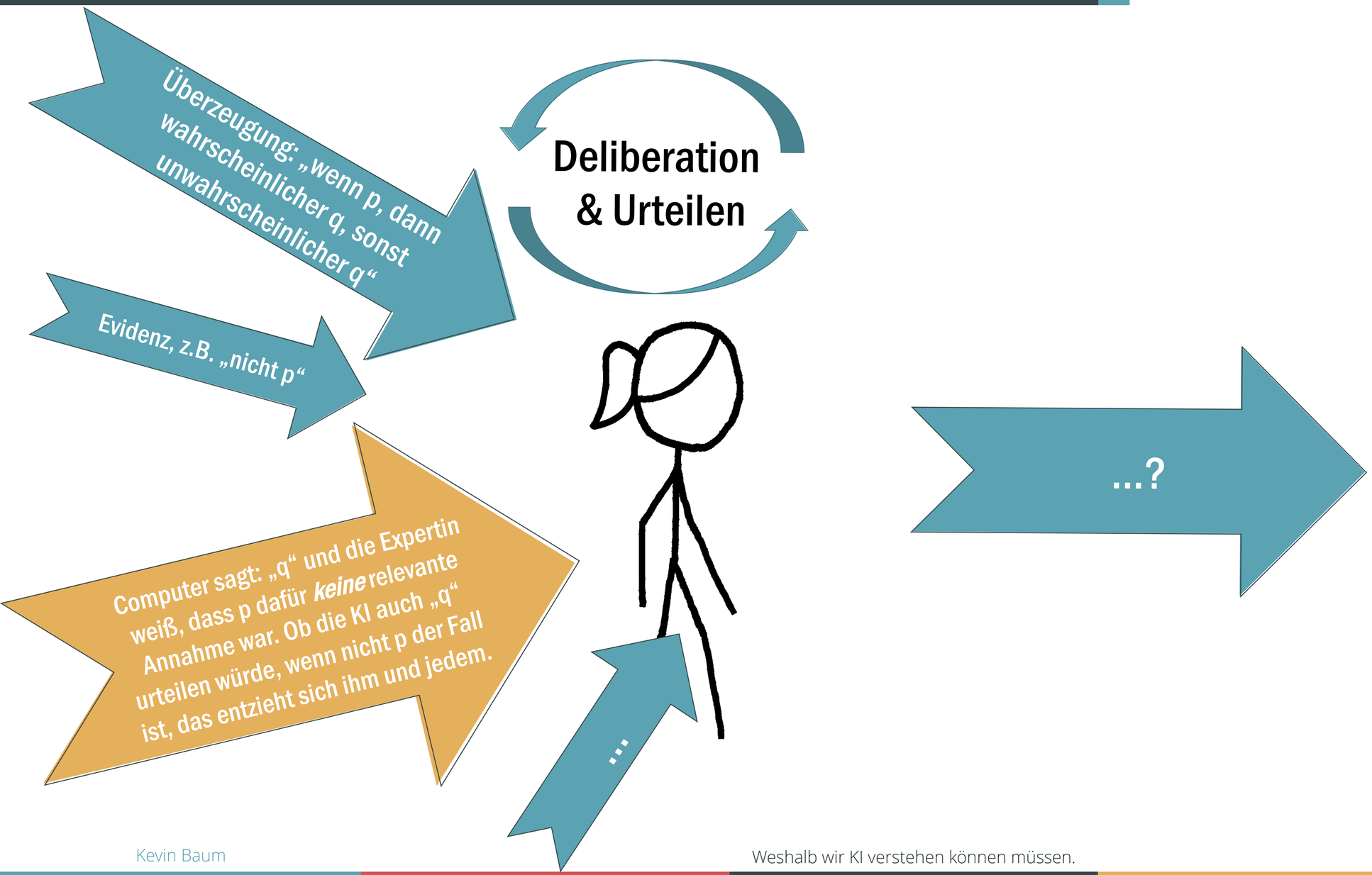



**Voraussetzung für Verantwortung, dafür, seinen Job machen zu können…**

Was tun im Falle sich widersprechender Evidenz oder neuer Informationen?

Reicht es, die Inputs der Black Box zu kennen?

# Das ganz-oder-gar-nicht-Problem

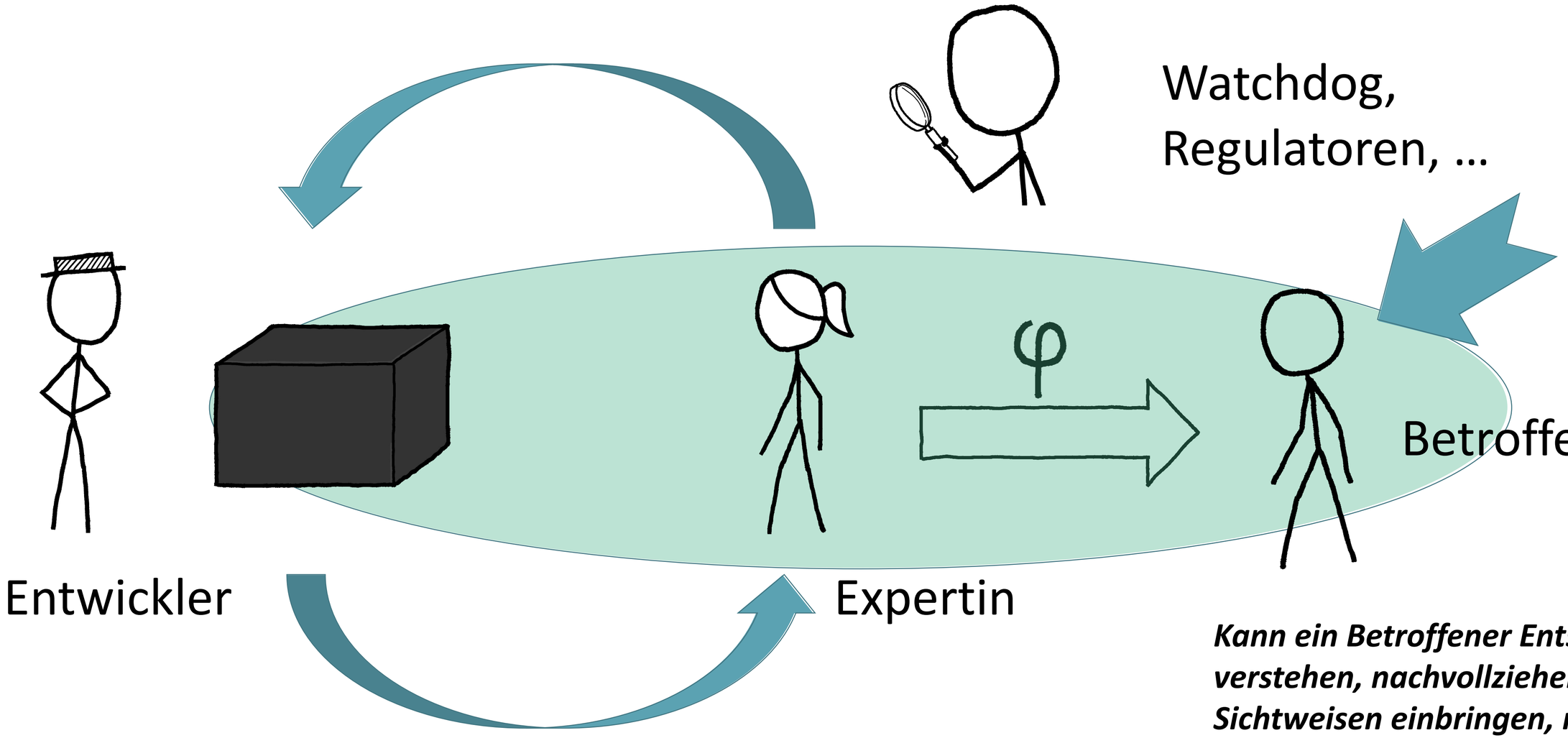



„Folge dem Computer oder eben nicht!“

Abwägen und angemessen, teilweise berücksichtigen, kann er nicht.

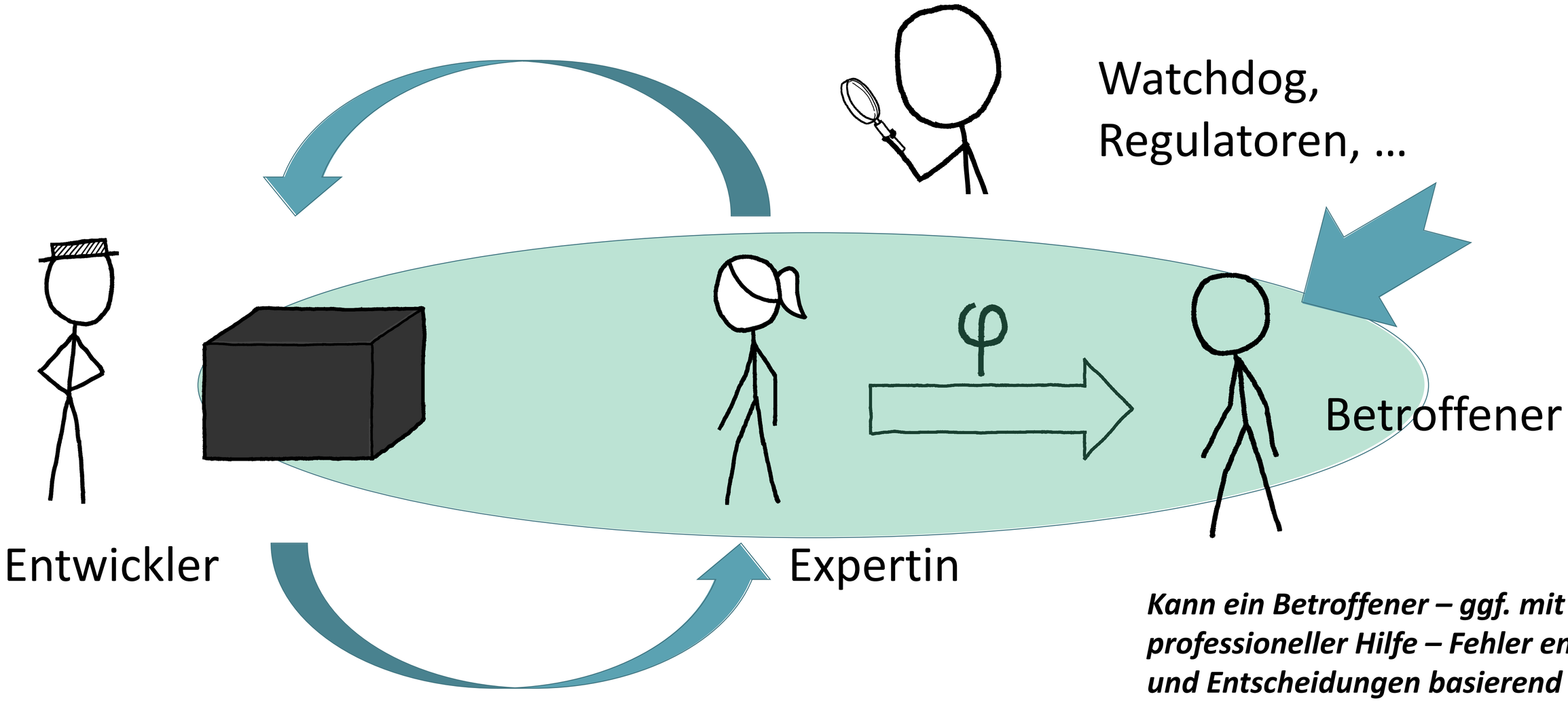
→ Willfähriger Entscheidungshelfer oder willkürlicher Verweigerer?

# Typ-3-Transparenz: Relevante Perspektiven




***Kann ein Betroffener Entscheidungen verstehen, nachvollziehen, eigene Sichtweisen einbringen, mögliche Fehler erkennen?***

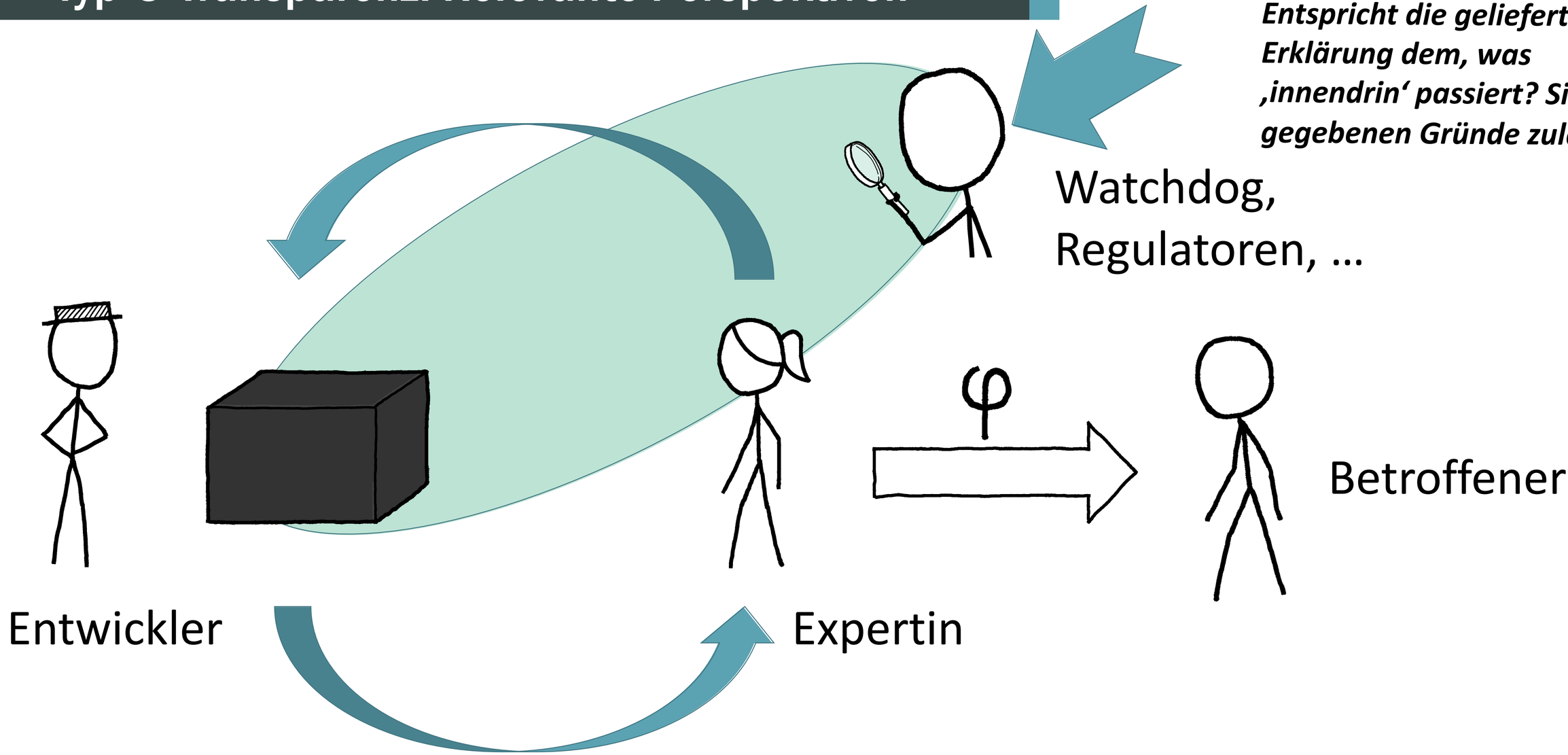
# Typ-3-Transparenz: Relevante Perspektiven




***Kann ein Betroffener – ggf. mit professioneller Hilfe – Fehler entdecken und Entscheidungen basierend auf den Erklärungen anfechten?***

# Typ-3-Transparenz: Relevante Perspektiven

***Im Zweifelsfall & Transparenz vom Typ 2 vorausgesetzt: Entspricht die gelieferte Erklärung dem, was ‚innendrin' passiert? Sind die gegebenen Gründe zulässig?***




# Typ 3:

# Transparenz des Entscheidungsprozesses & Zugang/Zugriff auf den internen, deliberativen Prozess

*Umsetzbarkeit?*

- Theoretisch u.U. sehr anspruchsvoll bis unmöglich (*Machine-Learning*)
- Viele begriffliche und empirische Fragen sind zu klären.
- Wahrscheinlich sind auch rechtliche Anpassungen nötig.


# Typ 3:

# Transparenz des Entscheidungsprozesses & Zugang/Zugriff auf den internen, deliberativen Prozess

*Warum wichtig?*

- Voraussetzung für *angemessene Reflektion des Experten* und damit für verantwortungsvolle Entscheidungen und angemessenen Umgang mit Urteilen
- Womöglich eine Voraussetzung für hinreichend *gerechte* Gesamtsysteme
- Voraussetzungen für *Einzelfallprüfung* (u.a. False Positives/Negatives erkennen)
- Rechtliche Anforderungen: Nicht bloß statistische Aussagen nötig.

# Wie können wir Erklärbarkeit ermöglichen und gewährleisten?

# Wir wissen es *noch* nicht …
# … aber wir arbeiten an der UdS dran!

Aktuelles
Uni-Porträt
Fakultäten|Einrichtungen
Wirtschaftsportal

Montag, 26. November 2018

https://www.uni-saarland.de/nc/universitaet/aktuell/artikel/nr/19923.html

**FOUNDATIONS OF PERSPICUOUS SOFTWARE SYSTEMS**

— Enabling Comprehension in a Cyber-Physical World —

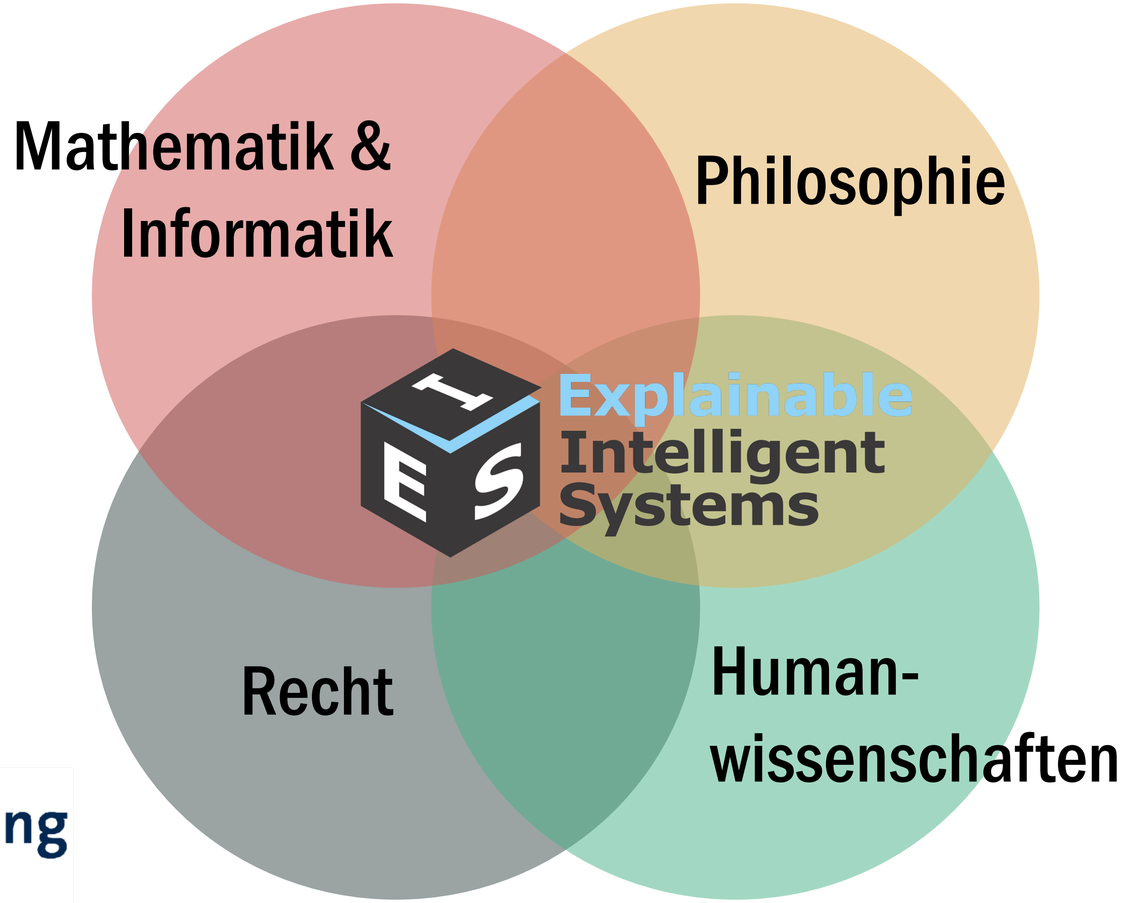

**Neuer Sonderforschungsbereich: Softwaresysteme sollen ihr Verhalten selbst erklären**

**Selbst Experten verstehen das Verhalten komplexer Softwaresysteme immer weniger. Dabei regeln diese inzwischen immer stärker unseren Alltag, sei es als intelligente Haussteuerung, im autonomen Fahrzeug oder in der industriellen Produktion. Wissenschaftler der Universität des Saarlandes, zweier Max-Planck-Institute und der Technischen Universität Dresden wollen jetzt in einem neuen Sonderforschungsbereich Mechanismen entwickeln, die nicht nur Experten, sondern auch Laien das Verhalten komplexer Softwaresysteme besser vermittelt. Die Deutsche Forschungsgemeinschaft fördert dieses Großprojekt mit elf Millionen Euro über vier Jahre hinweg.**

www.eis.science
www.explainable-intelligent.systems

# Transparenz durch Explainable AI

Erklärbare Künstliche Intelligenz (KI) = E**x**plainable **AI** = XAI




VolkswagenStiftung

# Transparenz: Mindestens drei Bedeutungen

Typ 1

~ Wissen, dass ein System über mich urteilt und auf Basis welcher Daten und was mit diesen geschieht

≠

Typ 2

~ Grobes Verstehen und formal nachvollziehbar machen

≠

≠

Typ 3

~ Nachvollziehbarkeit der individuellen Beurteilungen für Experten und Betroffene

# Offene Fragen: Mapping

**Welche Methode für welche Aufgabe?**

- Viele Eigenschaften der Methoden zur Erklärungsgenerierung:
  - Ausdrucksstärke
  - Algorithmische Komplexität…
- Viele Eigenschaften von Erklärungen:
  - Genauigkeit
  - Stabilität
  - Verständlichkeit…

# Offene Fragen: Mapping

**Welche Methode für welche Aufgabe?**

- **Welche Erklärungsmethode** für **welche Aufgabe/welchen Aspekt**?
- Wer **definiert Anforderungen** und wer **überprüft**, ob sie erfüllt sind, und **wie**?
- **Ziel** muss sein, die **Ansprüche**, die ein **Bürger** hat/haben soll, möglichst **klar** zu **formulieren**.
  Was soll der Effekt sein, was soll jemand damit tun können?
- Dann braucht es begleitende **Studien** (und ggf. neue Entwicklungen), um **herauszufinden, was welche Methode leisten kann.**

# Danke für Ihre Aufmerksamkeit!