Vertrauen in die digitale Welt. Wie kann diese verbessert werden?

Prof. Dr. Martin Hartmann

Universität Luzern

Seminar für Philosophie

Struktur des Vortrags

- 1: Thematische Einführung: Was ist Vertrauen?
- 2: Können wir digitalen Informationen vertrauen? Reicht es nicht, wenn sie verlässlich sind (trust vs. reliance)?
- 3: Beispiele: (a) Vertrauen in digitale Gesundheitstechnologien; (b) Vertrauen in digitale Informationen
- 4: Diskussion der Beispiele und eine These: Das medizinische Personal kann nicht substituiert werden!
- 5: Schluss

1: Thematische Einführung: Was ist Vertrauen?

- Vertrauen **zwischen Menschen**: **Akzeptierte Verletzlichkeit/Annahme des Wohlwollens des Vertrauensempfängers**
- ermöglicht uns, wichtige Ziele zu erreichen **(Verwirklichungsdimension)**
- erleichtert Kooperation **(Kooperationsdimension)**
- ist wertvoll an sich, zusätzlich zum Wert dessen, worum es im Vertrauen geht **(Wertdimension)**
- im Vertrauen erkennen wir einen anderen an als jemand, dem vertraut werden kann **(Anerkennungsdimension)**

1: Thematische Einführung: Was ist Vertrauen?

- ist riskant; im Vertrauen erhält der andere Spielräume, die wir nicht kontrollieren wollen (selbst wenn wir könnten) **(Risikodimension, Verletzbarkeitsdimension)**
- Vertrauen ist kein Ding, das wir schenken, es ist eine Art der Beziehung (eine Praxis), die von gemeinsamen Werten getragen ist **(Praxisdimension)** und die nicht erzwungen werden kann **(zum Vertrauen muss es Alternativen geben, sonst ist es kein Vertrauen)**
- Was macht jemanden vertrauenswürdig? Aufrichtigkeit, Ehrlichkeit, Kompetenz, Verlässlichkeit: **Welt und Wort passen zusammen** (Onora O'Neill) **(Dimension der Anhaltspunkte)**

1: Thematische Einführung: Was ist Vertrauen?

- Vertrauen und Sich-verlassen-auf (reliance): Im Vertrauen setzen wir auf das Wohlwollen des anderen, **das durch unser Vertrauen ausgelöst oder bestätigt wird**, wenn wir uns auf ihn verlassen, erwarten wir, dass er tut, wozu er etwa dienstlich verpflichtet ist (er muss uns nicht mögen und er kann einfach nur sanktionsfrei Geld verdienen wollen).

2: Können wir digitalen Informationen vertrauen? Reicht es nicht, wenn sie verlässlich sind (trust vs. reliance)?

- **Vertrauen zwischen Mensch und Technik, Mensch und digitalen Informationen** – kann es das geben?
- „The Trustworthy AI story is a marketing narrative invented by industry, a bedtime story for tomorrow's customers. The underlying guiding idea of a 'trustworthy AI' is, first and foremost, conceptual nonsense. **Machines are not trustworthy; only humans can be trustworthy (or untrustworthy).** If, in the future, an untrustworthy corporation or government behaves unethically and possesses good, robust AI technology, this will enable more effective unethical behaviour. Hence the Trustworthy AI narrative is, in reality, about developing future markets and using ethics debates as elegant public decorations for a large-scale investment strategy. At least that's the impression I am beginning to get after nine months of working on the guidelines." (Thomas Metzinger: **EU guidelines: Ethics washing made in Europe)**

2: Können wir digitalen Informationen vertrauen? Reicht es nicht, wenn sie verlässlich sind (trust vs. reliance)?

- **Lack of agency-These**: „Unlike the human clinician, AI systems have no goodwill towards us, not any motivation to act in our interests. … AI systems are not the appropiate objects of moral responsibility.“ (Joshua Hatherley, Limits of Trust in medical AI) – „Only a human being capable of autonomy can be worthy of active reflexive trust, because trustworthiness implies **accountability**.“ (Bjorn Myskja et al., Personalized medicine, digital technology and trust: a Kantian account)
- **Reliability-These**: Wir können uns eventuell auf AI verlassen, aber schon aus begrifflichen Gründen können wir ihr nicht vertrauen. Vertrauen ist mehr als Sich-Verlassen-auf!

Also ist mein Vortrag hier zu Ende?

2: Können wir digitalen Informationen vertrauen? Reicht es nicht, wenn sie verlässlich sind (trust vs. reliance)?

- Nein. Vertrauen hängt an unseren **Einstellungen** zueinander und in Technik oder digitale Informationen. Die blosse Nutzung einer Technologie sollte nicht schon mit Vertrauen verwechselt werden. Wir übertragen viele Elemente zwischenmenschlichen Vertrauens auf Technik und Digitales. Wir **erweitern** also unseren Vertrauensbegriff, wir verwenden ihn nicht einfach nur falsch, wenn wir von Technikvertrauen oder digitalem Vertrauen reden. Die Frage ist nicht: Falsch oder richtig, sondern: beschreibt der erweiterte Vertrauensbegriff unsere Praxis angemessen, halten wir ihn für legitim? **Die Antworten auf diese Fragen sind offen und nicht schon geklärt. Denn: Die Erweiterung mag mit einer Veränderung des Vertrauensbegriffs einhergehen, die bislang zentrale Elemente in den Hintergrund rückt.**
- Beispiel von Andreas Gutzeit (Der freundliche Halbautomat)

3: Beispiele: (a) Vertrauen in digitale Gesundheitstechnologien; (b) Vertrauen in digitale Informationen

- „Gut, das wir das Gerät hier hatten und es der Patientin geholfen hat, oder Herr Doktor? Sie, die schon fast ein ganzes Jahrhundert erlebt hat, scheint spontan eine **persönliche Beziehung** zum AED als Halbautomaten entwickelt zu haben. … Beginnt Künstliche Intelligenz damit, dass eine Neunzigjährige einem Halbautomaten mehr Vertrauen schenkt als ausgebildeten Ärzten?“
- Gutzeit wendet den Vertrauensbegriff an. Zu Recht? Vielleicht hält die alte Dame ihn, den Arzt, einfach nur für weniger verlässlich als das Gerät. Dass Gutzeit trotzdem von Vertrauen spricht, hängt an dem Element: persönliche Beziehung.

3: Beispiele: (a) Vertrauen in digitale Gesundheitstechnologien; (b) Vertrauen in digitale Informationen

- Und: Wenn wir von Vertrauen in Technik oder KI reden geht auch nach Gutzeit etwas verloren. Zum Beispiel ein Gespür für die **Fehlbarkeit von Technik und die Notwendigkeit menschlicher Übersetzung technischer Imperative in konkrete Praxis**. Wenn wir den Vertrauensbegriff erweitern, verlieren wir Elemente, die zwischenmenschliches Vertrauen tragen. **Wir nähern den Begriff an Verlässlichkeit an**.

3: Beispiele: (a) Vertrauen in digitale Gesundheitstechnologien; (b) Vertrauen in digitale Informationen

(a) Digitale Gesundheitstechnologien (wearables, apps zur Selbstdiagnose, Genomsequenzierung, mobile health, personalisierte Medizin). Hier auch eine typische Aussage, die zeigt, wie sehr Vertrauen an Verlässlichkeit angeglichen wird: „In terms of digital health technologies, we hypothesize that trust is likely to develop if the risks and uncertainties associated with their use **can be minimized**." (Adjekum et al., Elements of Trust in Digital Health Systems)

- Faktoren, die Vertrauen (in diesem Sinne) erhöhen: Nützlichkeit, Unkomplizierter Gebrauch, Datensicherheit, Empfehlungen von Familienmitgliedern, Schutz der Privatsphäre, Personalisierbarkeit, face-to-face-Kontakt vor Gebrauch (mit Arzt), Reputation der Anbieter. Besonders schwierig: Einschätzung der Qualität der Informationen!

3: Beispiele: (a) Vertrauen in digitale Gesundheitstechnologien; (b) Vertrauen in digitale Informationen

(b) Vertrauen in digitale Informationen

- Faktoren, die Vertrauen erhöhen: Erfahrung mit der Informationsquelle, Bestätigung der Qualität durch verschiedene Akteure oder multiple Quellen, emotionale Bindung (Design), Reputation, Zertifizierung einer Quelle durch unabhängige Organisationen, Übereinstimmung der Werte und Ziele einer Quelle oder ihrer Informationen mit den eigenen Werten und Zielen (resonate with style, arguments, or objectives presented in information), Genauigkeit, Objektivität, Gültigkeit der Informationen

4: Diskussion der Beispiele und eine These: Das medizinische Personal kann nicht substituiert werden!

- Problem: Wir vertrauen einer Information, wenn sie genau oder passend ist. Aber woher wissen wir das? In vielen Argumenten, die erläutern sollen, wann wir einer Information vertrauen, taucht eine **petitio principii** auf, wir müssen einer Quelle in gewisser Weise schon vertrauen, bevor wir bereit sind, die von ihr gelieferten Informationen für genau oder passend zu halten. Angesichts der Komplexität vieler Informationen: **Wie sollen wir hier überhaupt Vertrauen entwickeln?**
- Ein Hauptgrund dafür, Vertrauenswürdigkeit an Verlässlichkeit anzugleichen, **ohne diese Angleichung zu bemerken**, hat genau damit zu tun: Es ist leichter, sich auf jemanden zu verlassen als jemanden zu vertrauen. So vertrauen wir Technik nicht, weil sie funktioniert; wir vertrauen ihr, wenn sie funktioniert, aber für Vertrauen braucht es mehr als Funktionalität

4: Diskussion der Beispiele und eine These: Das medizinische Personal kann nicht substituiert werden!

- Was wir dann übersehen: Digitale Informationen legen sich nicht selbst aus, sie müssen übersetzt und verständlich gemacht werden; technische Geräte enthalten oft Anweisungen oder Verhaltensskripte, die problematisch sind, wenn sie undurchsichtig sind. Kurz: Je anspruchsvoller der Umgang mit digitalen Systemen ist, desto schwieriger wird es, Vertrauen intelligent zu schenken. Wir vertrauen oft blind (und gilt das nicht für alle stakeholder?) oder gar nicht. Wir sprechen von Vertrauen, verlassen uns aber oft auf Technik, ohne wirklich gut zu wissen, ob das immer sinnvoll ist.

4: Diskussion der Beispiele und eine These: Das medizinische Personal kann nicht substituiert werden!

- Und: Patienten werden wieder abhängiger von der Expertise des medizinischen Personals. Wir unterschreiben Dinge, die wir nicht verstehen. Heisst: Ein milder Paternalismus ist längst wieder da, allerdings entledigen wir uns seiner unangenehmen Seiten durch *informed consent*, der nicht wirklich *informed* ist. Nehmen Sie genetische Informationen. Myskja und Steinsbekk schreiben: „The genomic information will be more suitable as a diagnostic tool for the expert, due to the sheer amount of information in need of interpretation and validation.“ Richtig. Und dann schreiben sie: „But the time for medical paternalism is over.“ Was nicht sein darf, kann nicht sein.

4: Diskussion der Beispiele und eine These: Das medizinische Personal kann nicht substituiert werden!

- Was heisst all das für die Vertrauensfrage: Sie wird immer wichtiger, aber sie bleibt gebunden an Personen. Dass wir Personen durch KI oder Algorithmik ersetzen zu können glauben, beruht auf einem falschen Vertrauensbegriff. In Wirklichkeit brauchen wir weiterhin **menschliche Kontaktpunkte**, die uns Informationen übersetzen und zugänglich machen, die uns helfen, gute verlässliche Vertrauensquellen zu finden und die Verantwortung übernehmen, wenn Dinge schief gehen.
- Vergessen wir nicht die leibliche Dimension des Vertrauens. Wir sprechen oft von face-to-face; Vertrauen entsteht aber oft in der Anwesenheit im selben Raum, im Leib-zu-Leib.

4: Diskussion der Beispiele und eine These: Das medizinische Personal kann nicht substituiert werden!

- „Um einer Person zu vertrauen, so scheint es, muss ich mich in gleicher Weise ihr gegenüber verletzbar machen wie sie sich mir gegenüber verletzbar machen muss. Vertrauen beruht zum Teil darauf, dass der je andere keinen Vorteil aus dieser Verletzlichkeit zieht. Ich muss im selben Raum mit einer Person sein, muss wissen, dass sie mich physisch verletzen oder öffentlich erniedrigen könnte und wahrnehmen, dass sie das nicht tut, um ein Vertrauensgefühl zu entwickeln und mich dann in weiteren Hinsichten dieser Person gegenüber verletzbar zu machen.“ (Hubert Dreyfus, On the Internet)

5: Schluss

- Wir sollten nicht so sehr überlegen, wie wir digitalen Systemen besser vertrauen, wir sollten eher fragen, was Vertrauen in diesem Zusammenhang überhaupt heissen kann. Meine These wäre, dass Vertrauen in diesem Zusammenhang immer ein abgeleitetes Vertrauen ist: wir vertrauen digitalen Systemen in dem Masse, in dem wir den Personen vertrauen, die sie uns nahebringen oder mit uns verwenden. Damit werden diese Personen immer wichtiger, wichtiger vielleicht als ihnen lieb ist, denn es gibt durchaus Anreize, Verantwortung an die Apparate und Apps abzugeben. Wenn wir es mit Vertrauen zu tun haben wollen, wir das nicht gelingen.