Diskussionen

Was heißt hier autonom? Über die moralische Autorenschaft selbstfahrender Autos

Von Lukas Ohly

1. Einleitung

In Bei unausweichlichen Autounfällen mit Personenschäden entscheiden die Menschen hinter dem Steuer situativ, wem sie ausweichen und wen sie stattdessen treffen. Wer situationsgemäß entscheidet, verhält sich *zwingend* und wendet kein Lebenswerturteil an.[1]

Anders autonom fahrende Autos in solchen Situationen: Entweder muss eine Präferenz für bestimmte Menschengruppen einprogrammiert werden oder das Auto entscheidet nach einem Zufallsgenerator. In beiden Fällen wird die Menschenwürde verletzt. Das gilt auch für den Vorschlag einer »vorweg programmierbaren Schadensminderung … bei der die Identität der Verletzten oder Getöteten … noch nicht feststeht.«[2] Soll nämlich »die Programmierung das Risiko eines jeden einzelnen Verkehrsteilnehmers in gleichem Maße« reduzieren[3], so funktioniert diese Gleichheit nur, wenn sie *zufällig* aufgehoben werden *dürfte*. Die Software würde dabei auf der Unterstellung basieren, der Wert des menschlichen Lebens sei *zufälligerweise* einem anderen vorzuziehen. Hier deutet sich ein Unterschied zum situativ zwingenden Eingriff des autonomen Verhaltens eines menschlichen Fahrers an. Solche Veränderungen des Autonomie-Phänomens sind das Thema dieses Beitrags. Schwerpunkt ist dabei das Verhältnis von Entscheidungen zur Instanz, die sie trifft.

2. Warum Künstliche intelligente Systeme keine Subjekte sind[4]

Bislang kennen wir nur moralische Agenten, die auch Subjekte sind. Subjektivität ist dabei nichts Gemachtes, sondern ein *Widerfahrnis*.[5] Denn wenn ich das mir selbst hätte vornehmen können, ein Erleben zu haben, müsste ich ja vorher schon da gewesen sein. Das führt in einen logisch unendlichen Regress.

1. *Johannes Fischer*: Verstehen statt Begründen. Warum es in der Ethik um mehr als nur um Handlungen geht; Stuttgart 2012, 175.
2. *Ethik-Kommission Automatisiertes und Vernetztes Fahren*. Eingesetzt durch den Bundesminister für Verkehr und digitale Infrastruktur, Juni 2017, 18.
3. Ebd.
4. *Lukas Ohly*: Schöpfungstheologie und Schöpfungsethik im biotechnologischen Zeitalter; Berlin 2015, 148–153.
5. Ebd., 185.

Zeitschrift für Evangelische Ethik, 63. Jg., S. 294 – 300, ISSN 0044-2674

Es kann aber auch kein Anderer mein Ich herstellen. Denn ein Anderer hat ein eigenes Ich und kann nicht in mein Ich springen.[6] Mein Ich ist für jeden Anderen transzendent.[7] Andere können höchstens objektive Bedingungen herstellen – z.B. ein Gehirn nachbauen – in der Hoffnung, dass daraus ein Ich entsteht. Aber wenn dann ein Ich entsteht, wird es nicht durch die objektiven Bedingungen hinreichend bestimmt.[8] Denn der Sprung von objektiven Bedingungen zu einem Subjekt kann nicht objektiv vorgenommen werden.

Nun könnte zwar wie durch ein Wunder ein Computer ein Subjekt werden. Wenn das der Fall ist, dann aber nicht, weil er gemacht ist. Doch was sollte uns dann überhaupt zur Idee veranlassen, ein Computer sei ein Subjekt? In der Interaktion mit reaktionsfähigen Systemen macht es einen erheblichen Unterschied, ob wir es mit einem gemachten Gegenüber zu tun haben oder mit einem gewordenen Subjekt. Wir schämen uns z.B. nicht vor einem Computer. Und selbst, wenn wir es tun, nehmen wir ihn immer auch als Objekt wahr, vor dem man sich nicht schämt.[9] Während wir uns zudem vor Subjekten schämen, auch wenn sie abwesend sind und wenn der Anlass unserer Scham zeitlich zurückliegt, erlischt die Scham vor Computern spätestens mit ihrem Abwesendsein.[10] Unsere Ahnung, der Computer könnte ein Subjekt sein, wird also immer zugleich neutralisiert, weil wir ihn im Modus des Ein- und Ausschaltens wahrnehmen und seinen Blicken damit prinzipiell entkommen können.[11] Deshalb fehlt der Vorstellung, ein Computer könne ein Subjekt sein, die phänomenologische Adäquanz.

Im Widerfahrenscharakter besteht der theologische Charakter der Subjektivität. Darauf wird zurückzukommen sein.

3. Die Autonomie des Agenten und seine Identität

Nun stellt ohnehin kein Autobauer den Anspruch, selbstfahrende Autos als Subjekte herzustellen. Intelligente Maschinen werden zunehmend autonom, auch ohne Subjekte zu sein.[12] Allerdings setzt unser Verständnis von Autonomie die Identität eines Agenten voraus. Wenn eine Maschine nach einem Programm läuft, so folgt sie den heteronomen Regeln eines Programmierers. Um autonom zu sein, muss die Maschine die Regeln, nach denen sie abläuft, selbst für sich gesetzt haben. Nach welchen Regeln agiert aber ein solches selbstlernendes System, wenn sich sein Verhalten nicht vorhersagen lässt? Diese Regeln liegen dann im Verborgenen. Selbst wenn jemand das Verhalten eines selbst fahrenden Autos beobachtet und die Regeln im

6. *Dieter Birnbacher*: Künstliches Bewußtsein; in: Thomas Metzinger (Hg.): Bewußtsein. Beiträge aus der Gegenwartsphilosophie; Paderborn 2005[5], 713–729, 728.
7. *Dietrich Bonhoeffer*: Sanctorum Communio. Dogmatische Untersuchung zur Soziologie der Kirche; München 1986 (DBW 1), 32.
8. Ohly (wie Anm. 4), 117.
9. *Christopher Scholtz*: Alltag mit künstlichen Wesen. Theologische Implikationen eines Lebens mit subjektsimulierenden Maschinen am Beispiel des Unterhaltungsroboters Aibo; Göttingen 2008, 285. *Lukas Ohly*/*Catharina Wellhöfer*: Ethik im Cyberspace; Frankfurt 2017, 288f.
10. Ohly (wie Anm. 4), 128f.
11. Ohly/Wellhöfer (wie Anm. 9), 262.
12. *Bernhard Irrgang*: Posthumanes Menschsein? Künstliche Intelligenz, Cyberspace, Roboter, Cyborgs und Designer-Menschen – Anthropologie des künstlichen Menschen im 21. Jahrhundert; Wiesbaden/Stuttgart 2005, 140.

Nachhinein rekonstruiert, kann er die Rekonstruktion nicht für Vorhersagen nutzen. Dasselbe ist der Fall, wenn die algorithmischen Muster der bisherigen Prozesse rekonstruiert werden: Daraus lassen sich keine sicheren Vorhersagen für künftige Fahrstrategien des Autos treffen.[13] Als selbstlernendes System agiert es allein nach internen Regeln.

Aber was bedeuten »interne Regeln«, wenn der Computer kein Subjekt ist, das sie sich gibt? Wir gestehen Menschen zu, dass sie nach internen Regeln agieren: Wir können von ihnen durch ihre Fahrweise überrascht werden, und wenn wir sie danach fragen, lassen wir die Antwort gelten, dass sie keine anderen Gründe hatten als diesmal einfach anders fahren zu wollen.[14] Was bedeutet aber Autonomie, wenn das System, das sich seine internen Regeln »selbst« gibt, gar kein »Selbst« ist? Seine »internen Regeln« sind dann weder objektiv eindeutig, noch sind sie subjektiv. Daraus folgt, dass man dann nicht von Regeln sprechen kann. Zwar könnte man einräumen, dass ein selbstlernendes System nach der Regel der Optimierung agiert. Aber was Optimierung für den Computer bedeutet, verändert sich durch jede Optimierungsstrategie, die er als selbstlernendes System entwickelt. So wird sich ein selbstlernendes Auto danach optimieren, möglichst gut zu fahren. Aber was möglichst gut ist, hat keine interne Regel. Ebenso wenig kann ein Beobachter ein- für allemal eine solche Regel der Fahrweise des Autos entnehmen. Ein selbstlernendes System kann mit den Parametern spielen und manche Verschlechterungen für die Gesamtoptimierung nutzen: Geht es z.B. darum, möglichst sicher zu fahren, muss die Optimierung des Ziels, möglichst schnell anzukommen, zurückgestellt werden. Ist die Koordination mit anderen Verkehrsteilnehmern zu optimieren, kann das dazu führen, dass sich der Energieverbrauch verschlechtert. Die Fahrweise kann sich durch das Selbstlernen permanent verändern, ohne dass wir sichere Aussagen darüber treffen können, wie sich der Fahrstil sogar im Lauf derselben Fahrt noch ändern wird. Selbst der Computer wäre zu sicheren Vorhersagen nicht in der Lage, da er sich jederzeit revidieren könnte. Denn es fehlt ihm Identität. Es ist zwar dasselbe Auto, aber seine Autonomie liegt gerade nicht in seiner objektiven Identität oder an den objektiven Eigenschaften, sondern im Selbstlernen und internen Agieren. Und genau hier fehlt eine Instanz, die »selbst« lernt.[15]

---

13. *Anja Dahlmann*: Militärische Robotik als Herausforderung für das Verhältnis von menschlicher Kontrolle und maschineller Autonomie; ZEE 61/2017, 171–183, 182.
14. *Ludwig Wittgenstein*: Philosophische Untersuchungen § 217. Nun könnte man gerade mit Wittgenstein einwenden, dass es keine privaten Sprachen und somit keine internen Regeln von Subjekten gibt. Somit müssen »interne Regeln« etwas Öffentliches haben. Das ist aber tatsächlich der Fall: Menschen rechnen einander intersubjektiv die interne Regeln zu, dass sich das Widerfahren von Subjektivität in lebensgeschichtlichen Entscheidungen niederschlägt. Oder anders: Das Widerfahren von Subjektivität bildet seine subjektive Identität. S. *Lukas Ohly*: Können wir autonom unser Gehirn manipulieren, bis wir jemand anderes sind? Zum Verhältnis von Neuroethik, Bewusstseinsphilosophie und Theologie; NZSTh 56/2014, 141–159, 150, 156f.
15. Eilert Herms hat für die »Identität des Werdens nur diejenigen Bedingungen des jeweiligen Werdens« zugelassen, »die in der Dauer von Gegenwart alle Übergänge, die zu dem jeweils identischen bestimmten Werden gehören, überdauern« (*Eilert Herms*: Systematische Theologie Bd. 2; Tübingen 2017, 1189). Herms hat diese Bedingungen nur in Instanzen gefunden, die selberwirken (1190). Könnte diese Charakterisierung auch für die Identität autonomer künstlicher Systeme gelten? Da Herms' Charakterisierung eine Beschreibung immer nur zeitlich rückwirkend ermöglicht, kann man nicht wirklich von einer »Identität des Werdens« sprechen, sondern nur von einer »Identität des Gewordenen«. Personen wiederum wären als identisches Werden damit überfordert, alle ihre Übergänge selberwirkend zu antizipieren, die zu ihnen gehören. Somit ist Herms' Kriterium weder theoretisch hinreichend für die Bestimmung der Identität des Werdens noch anwendbar für die Identität autonomer künstlicher Systeme.

4. Der theologische Charakter der Autonomie

Der Widerfahrenscharakter der Subjektivität[16] ist nach Schleiermacher theologischer Art[17], weil nur *eine* Instanz für mein Ich bürgen kann, und zwar eine Instanz *außerhalb der Welt* (Gott).[18] Der schlechthinnigen Abhängigkeit entspricht, dass es *genau eine Instanz* gibt, die schlechthinnig frei ist[19] – und dass damit alles, was es gibt, auch von dieser Instanz abhängig ist.[20] Gott wiederum findet seine Bestimmung ausschließlich im Selbstvollzug: »Ich werde sein, der ich sein werde« (Ex. 3, 14).

Das trifft auch auf menschliche Autonomie zu – nur mit dem Unterschied, dass menschliche Autonomie von Gott schlechthinnig abhängig ist. Das menschliche Ich, das autonom agiert, widerfährt sich zugleich – unausweichlich fremdbestimmt. Der Mensch könnte nur sagen: »Ich werde sein, der ich sein werde, *sofern ich sein werde*.« Die *Verlässlichkeit* der menschlichen Ich-Identität verdankt sich demnach nicht sich selbst.

Autonome künstliche Systeme nun haben keine interne Instanz, die interne Regeln formuliert. Allerdings simulieren sie eine schlechthinnige Freiheit, weil ihr Verhalten keine erwartbare Verlässlichkeit bietet. Für sie gilt in Abwandlung der göttlichen Selbstvorstellung: »*Es wird sein, was sein wird*.«

Damit tritt in unserer Interaktion mit einem selbstfahrenden Auto der Widerfahrenscharakter hervor. Noch stärker als in der subjektiven Selbsterfahrung tritt das Widerfahren seiner Autonomie hervor, gerade weil es keine Instanz ist, die eine Entscheidung trifft. Daher kann es als quasi-göttlich wahrgenommen werden. Zwar ist das selbstfahrende Auto davon abhängig, dass es ein Auto ist – mit allen Konsequenzen der schlechthinnigen Abhängigkeit. Als selbstlernendes System trifft es aber seine Entscheidungen unabhängig davon, dass es ein Auto ist.

Es ist daher damit zu rechnen, dass sich Menschen an selbstfahrende Autos ausgeliefert fühlen, eben weil der Widerfahrenscharakter ihrer Entscheidungen hervortritt: Als Verkehrsteilnehmer dürfte man das Verhalten des Autos immer wieder als Schicksalsschlag erleben. Und als Insasse gibt man sein Schicksal ganz in die Hände des Autos. Das ist eine andere Erfahrung als sich Subjekten anzuvertrauen, sich etwa als Patient in die Hände von Ärzten zu begeben. Mit Ärzten können Patienten Verträge schließen, weil Ärzte identische Träger ihrer Entscheidungen und damit zurechnungsfähig sind.

Ohne identische Instanzen zu sein, werden autonom fahrende Autos zum unbestimmten Agenten. Autonom sind nur ihre Entscheidungen. Genau darin besteht das Gefühl des Ausgeliefertseins und der Schicksalhaftigkeit. Um sie herum treten gezielte Ereignisse ein, ohne dass es jemanden gibt, der sie anzielt.

16. *Bernhard Waldenfels*: Phänomenologie der Aufmerksamkeit; Frankfurt a.M. 2004, 188.
17. *Friedrich Schleiermacher*: Der christliche Glaube nach den Grundsätzen der evangelischen Kirche im Zusammenhange dargestellt; Zweite Ausgabe (1831) Bd. 1; Berlin 1960, 28.
18. Deshalb kann diese Instanz auch nicht der Zufall sein. Denn der Zufall ist nicht verlässlich.
19. Schleiermacher (wie Anm. 17), 27.
20. Ebd.

5. Begrenzung des Schicksals im Straßenverkehr

Wenn wir selbstfahrende Autos im Straßenverkehr zulassen, so lassen wir dort Begegnungen zu, in denen wir dem theologischen Widerfahrenscharakter ihrer Autonomie ausgesetzt sind, ohne dass sich daraus verlässliche Gehalte erschließen lassen. Das bringt unsere bisherigen ethischen Sozialbezüge des Straßenverkehrs durcheinander. Der Straßenverkehr ist nämlich dadurch ausgezeichnet, dass er unsere Umweltbeziehung möglichst stark auf zwischenmenschliche Interaktion konzentriert. Anstelle schicksalhaft erlebter Widerfahrnisse wird der Fokus auf verlässliche Interaktionsgehalte von Menschen mit ihresgleichen gelegt. Der Straßenverkehr ist so organisiert, dass seine Teilnehmer möglichst wenig dem Widerfahrenscharakter des Unvorhergesehenen ausgeliefert werden.

Dadurch ist der Mensch die allererste Gefahrenquelle für den Straßenverkehr. Menschen treffen autonome Entscheidungen, die sie und andere in Gefahr bringen und die sich nicht kontrollieren lassen, weil auch autonome Entscheidungen von Menschen nach internen Regeln getroffen werden. Der entscheidende Unterschied zwischen zwischenmenschlichen und nicht-menschlichen Interaktionen im Straßenverkehr liegt aber darin, dass Menschen *Reziprozitätserwartungen* aneinander richten. Wir rechnen mit autonomen Entscheidungen bei uns selbst ebenso wie bei anderen, weil wir uns reziprok als Subjekte anerkennen – und weil wir reziprok anerkennen, dass wir nicht anders können als autonom zu sein (»Ich werde sein, der ich sein werde, *sofern ich sein werde*.«) Wir können daher um die Gefahr wissen, die wir für andere sind, ebenso wie wir um die Gefahr der anderen Verkehrsteilnehmer für uns wissen. Das bereits minimiert das Risiko immerhin. Die Gefahr, die Menschen füreinander sind, liegt auf derselben Ebene, nämlich des autonomen Subjekts. Dadurch kann man einigermaßen einschätzen, wie Subjekte sich verhalten, weil sie nämlich identische Instanzen sind.

Diese Reziprozitätserwartung fehlt nun gegenüber selbstfahrenden Autos. Zwar erwarten wir auch bei ihnen autonome Entscheidungen. Aber da diese Entscheidungen keiner Instanz zuzurechnen sind, liegen sie auf einer anderen Ebene. Selbstfahrende Autos können erwartbar in bestimmten Verkehrssituationen eine Gefahr für den Menschen darstellen, aber nicht weil sie erwartbar unerwartet agieren *wie wir*, sondern erwartbar unerwartet wie ein plötzliches Ereignis *ex nihilo*. Mit dieser Technik die Zahl der Verkehrstoten zu reduzieren, ist damit erkauft, dass der Straßenverkehr unheimlicher wird, weil man mit Schattenexistenzen zusammen fährt, die jederzeit unvorhersehbar reagieren können, ohne dabei Instanzen zu sein, die diese Reaktionen autonom ausführen.

Um die Reziprozitätserwartungen zu sichern, gibt es nur zwei Optionen:

1. Entweder bleiben selbstfahrende Autos vom Straßenverkehr ausgeschlossen.
2. Oder Menschen werden als Interaktionsteilnehmer im Straßenverkehr ausgeschlossen, indem der Verkehr vollautomatisiert wird und alle Agenten im Straßenverkehr selbstfahrende künstliche Systeme sind.

Die zweite Option müsste so konsequent umgesetzt werden, dass jeglicher systemische Einfluss des Menschen auf den Straßenverkehr ausgeschlossen ist. Der Mensch kann dann weder zwischendurch ins Lenkrad greifen noch auf die Fahrt einwirken, wenn ihm etwa aufgrund der Fahrweise des Autos schlecht wird. Das ist eine unattraktive Option. Deshalb ist für die erste Option zu votieren. Mit quasi-göttlichen Phänomenen im Straßenverkehr zu rechnen, gefährdet die menschliche Freiheit.

Gegen dieses Ergebnis könnte man einwenden, dass die Argumente eine zweifelhafte Analogie voraussetzen, nämlich dass selbstfahrende Autos ebenso *Fahrer* sind wie Menschen und dass menschliche Autofahrer mit ihnen *interagieren*. Tatsächlich könnte man solche Fahrzeuge aber auch anders interpretieren, nämlich etwa als Umweltbedingungen des zwischenmenschlichen Straßenverkehrs: Ebenso wie Menschen im Straßenverkehr mit automatischen Ampeln konfrontiert sind, ohne sie als sozialen Interaktionspartner anzuerkennen, wären autonome Autos automatische Abläufe, die Menschen zwar berücksichtigen müssen, ohne aber mit ihnen zu interagieren. Dementsprechend müsste der Einwand behaupten, dass autonom fahrende Autos *keine* Entscheidungen treffen – *und damit auch nicht autonom sind*. Das Problemverhältnis von Mensch und Auto würde also in meinem Beitrag auf einer inadäquaten Interpretation beruhen.

Die Pointe meiner Argumentation beruht aber gerade darin, dass sie diese Inadäquanz *im Ergebnis* einräumen kann, indem autonom fahrende Autos gerade nicht als Entscheidungs*instanzen* anerkannt werden, sondern als nicht-sozial eingefangene Schicksalsschläge. Wer also methodisch-begrifflich einen anderen Ausgangspunkt nimmt, würde das Ergebnis des vorliegenden Beitrags bereits in den Voraussetzungen vorwegnehmen: Autonom fahrende Autos fahren *irgendwie* – ohne dass sie dafür sozial stabile Erwartungen hervorrufen können. Vielmehr begegnen sie uns als quasi-göttliche Schicksalsschläge – anders als fest installierte und programmierte Ampeln.

6. Ergebnis und Konsequenzen

Menschen müssen mit unvorhergesehenen Ereignissen leben. Dabei thematisieren sie oft den Widerfahrenscharakter solcher Ereignisse, der sich auf derselben kategorialen Ebene befindet, auf der religiöse Menschen von Gott reden.[21]

Mit Unvorhergesehenem lässt sich leichter leben, wenn der Widerfahrenscharakter menschliche Züge hat, die verlässlich sind. Das begrenzt das Schicksalhafte und wahrt unsere Freiheit. Darin sehe ich den entscheidenden Beitrag des Christusgeschehens für die Ethik.[22] Denn im Christusgeschehen hat sich die Offenbarung Gottes an einen Menschen gebunden: »Ich und der Vater sind eins« (Joh. 10,30) beschreibt zwar keine Identität, aber eine eindeutige Zuordnung zwischen Gott und Mensch in Christus. Das »Ich werde sein, der ich sein werde« erhält somit eine irdische Lokalisierung. Bezogen auf die Begrifflichkeit, die ich in diesem Beitrag vorgeschlagen habe, heißt das: Der Widerfahrenscharakter von Ereignissen bindet sich an die Instanz eines Menschen, die die Identität des Widerfahrens repräsentiert. Die Zurechenbarkeit autonomer Akte eines Menschen hängt daher an zwei Bedingungen:

1. Das Subjekt muss dabei identisch sein.
2. Es verdankt seine Identität dem Widerfahrenscharakter seiner Subjektivität.

Insofern ist die Christologie Paradigma für die Zurechenbarkeit autonomer menschlicher Akte.

Die ethische Zukunft des Autos liegt darin, Assistenzfunktion für den Menschen zu übernehmen. Das intelligente Auto wird den Menschen beraten, warnen, seine je aktuelle Fahrtaug-

21. *Lukas Ohly*: Warum Menschen von Gott reden. Modelle der Gotteserfahrung; Stuttgart 2011, 24ff.
22. *Lukas Ohly*: Was Jesus mit uns verbindet. Eine Christologie; Leipzig 2013, 207f.

lichkeit überprüfen – es wird aber nicht in seine Entscheidungen eingreifen, es sei denn, diese Reaktionen sind einprogrammiert und keine autonomen Entscheidungen eines selbstlernenden Systems. Solche einprogrammierten Reaktionen – etwa einen Auffahrunfall zu vermeiden, den ein unkonzentrierter Fahrer ansonsten begehen würde – sind ethisch unproblematisch, wenn sie keine Wertentscheidungen über verschiedene Menschenleben treffen und aus dem Verkehrssystem nicht ausbrechen.[23]

Apl. Prof. Dr. Lukas Ohly, M.A. phil.
Goethe-Universität Frankfurt am Main
Fachbereich Evangelische Theologie
Systematische Theologie
Norbert-Wollheim-Platz 1
D–60323 Frankfurt am Main
ohly@kirche-ostheim.de

23. Ich danke Catharina Wellhöfer-Schlüter für ihre inspirativen Anmerkungen.