Autonoblog
Autonomes Fahren zwischen Hype und Wirklichkeit

Ethik & autonomes Fahren IV: Wie autonome Fahrzeuge wirklich entscheiden

Analyse, Ethik & autonomes Fahren, Wiki / Von David Knollmann

Im letzten Beitrag dieser Reihe haben wir gelernt, dass sich die Entscheidungssituationen in Trolley-Problemen von denen in der „real world" regelmäßig fundamental unterscheiden. Nun beurteilt Ethik Handlungen. Um zu einem Urteil über die Ethik autonomer Fahrzeuge zu gelangen, muss man sich also in einem nächsten Schritt genau anschauen, wie die „intelligente" Maschine im Fahrzeug *handelt* – wie sie *entscheidet*.

Sogleich wird man feststellen, dass

a) die Maschine völlig anders entscheidet als ein Mensch und mithin auch völlig anders als ein Mensch, der sich einer Dilemma-Situation gegenüber sieht und

b) dass sich aufgrund dieses Entscheidungsdesigns ganz andere Arten von ethischen wie sozialen Dilemmata ergeben.

Autonomes Fahren als Entscheidungsproblem: partially observable Markov-Decision-Problems (POMDP)

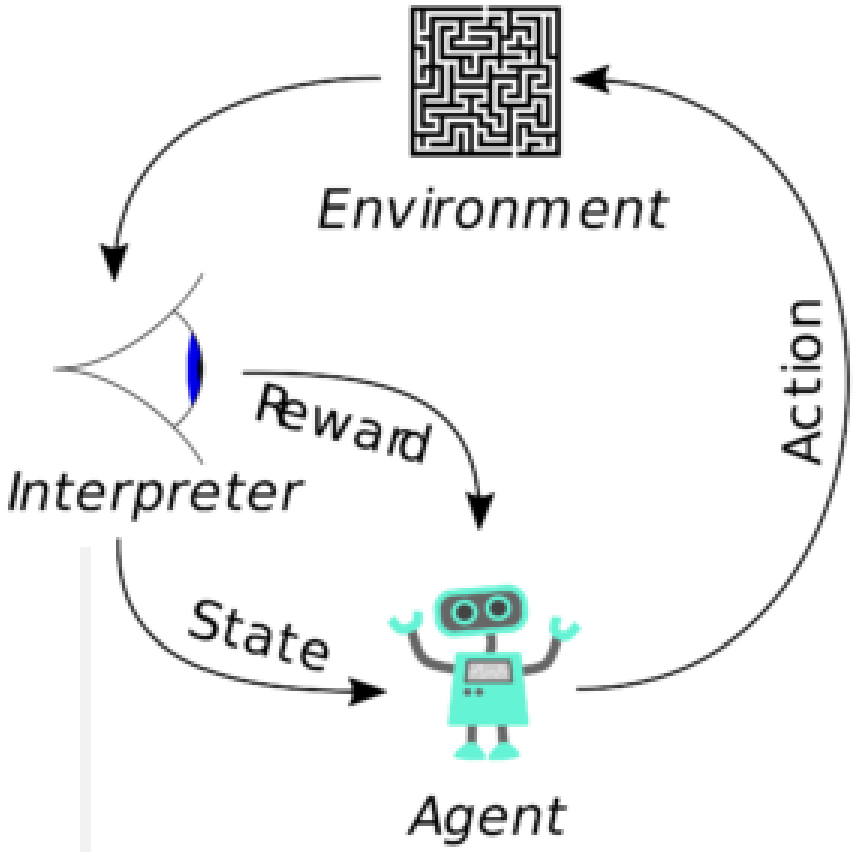
Ein mathematisches „Framework" des maschinellen Lernens (machine learning, ML), das in autonomen Fahrzeugen zum Einsatz kommt, ist das „partially observable Markov-Decision-Problem" (POMDP). Das Markowsche Entscheidungsproblem (MDP**,** Markov decision process) ist ein mathematisches Modell, das nach einer optimalen Entscheidung bzw. Entscheidungsfolge in einer sogenannten zugänglichen, indeterministischen Umgebung sucht. Die Entscheidungsfolge soll dabei den Nutzen des Handelnden („*agent*") maximieren. Eine Nutzendimension könnte etwa sein, möglichst schnell ans Ziel zu kommen oder Kollisionen zu vermeiden. Wie wir zuvor gelernt haben, unterscheidet jener Indeterminismus die reale Welt vom Gedankenexperiment des Trolley-Problems: Wo bei letzterem alle Konsequenzen der Entscheidung im Vorhinein bekannt sind, regiert in ersterer die Unsicherheit. Ein „partially observable" MDP (POMDP) liegt dann vor, wenn die Umgebung *nicht* vollständig zugänglich ist, weil diese aufgrund der begrenzten kognitiven Ressourcen der Maschine nicht zu einhundert Prozent erfasst werden kann: Die Sensortechnik im automonen Fahrzeug funktioniert nicht immer vollständig zuverlässig, sie hat etwa Reichweitengrenzen, kann getäuscht oder ihre Sicht verdeckt werden und

schließlich kann sich in der Zeit zwischen dem Eingang der Sensorsignale, ihrer Verarbeitung und Interpretation (*„latency"*), die Umgebung wieder verändert haben. Der *agent* (= die Maschine) muss deshalb ständig *Annahmen* über die möglichen sogenannten Zustände (*„states"*) in der Umgebung treffen und also *Wahrscheinlichkeitswerte* berechnen, wobei bestimmte Handlungen neue Informationen über die Zustände mit wiederum variierenden Wahrscheinlichkeitswerten liefern können. Zu jedem Zeitpunkt muss die Maschine eine Entscheidung treffen abhängig von ihrem aktuellen Zustand, ggf. dem vorangegangenen Zustand und dem zu erwartenden zukünftigen Zustand. Das POMDP ist also ein Modell, um *sequentielle* Entscheidungsprozesse in der realen Welt abzubilden, die auf *probabilistischen* Annahmen beruhen und mit *begrenzten kognitiven Ressourcen* umgehen müssen. Während also in Trolley-Problemen immer von *einer* gewichtigen Entscheidung zu *einem* Zeitpunkt die Rede ist, werden autonome Fahrzeuge immer zahlreiche *Entscheidungsfolgen* auf einer kontinuierlichen Zeitachse unter Unsicherheit bewerten müssen.

Die Maschine trainieren: reinforcement learning

Wie soll sich die Maschine aber nun konkret verhalten? Das POMDP gibt gewissermaßen den formalen Kontext vor für den Einsatz verschiedener ML-Konzepte. In autonomen Fahrzeugen kommen unter anderem Varianten des sogenannten *Reinforcement Learnings* zum Einsatz, das den *agent* durch den Einsatz von Belohnungen durch ein unbekanntes Umfeld *leitet* (=*supervised* learning). Dieses *bestärkende Lernen* funktioniert also durch eine Feedbackschleife, einen *Interpreter*, der den *agent* über etwaige Belohnungen und Bestrafungen sowie den neuen Zustand infolge der gewählten Aktion informiert. Die Maschine lernt also sequentiell durch *trial & error* über den Nutzen der gewählten Aktionen: Belohnung winkt da, wo die Aufgabe besonders gut erledigt wird, z.B. indem sicher und ohne Kollision von A nach B navigiert wurde. Weil Informationen über vergangene Feedbacks und korrespondierende Wahrscheinlichkeitsfunktionen im Maschinengedächtnis gespeichert werden, bildet sich idealerweise über Zeit und unzählige Wiederholungen der Feedbackschleifen ein *Verständnis* darüber aus, was optimale Entscheidungen bzw. Entscheidungsfolgen in zahlreichen unterschiedlichen, mit Unsicherheit behafteten Situationen sein können.

Die Herausforderung für den Informatiker, der die Maschine anleitet, liegt nun vor allem darin, die Ziele für die Maschine in dem ML-Modell so zu definieren, dass diese lernt, welche Aktionen mit den richtigen Beobachtungen der Umgebung korrespondieren – trotz bestehender Unsicherheiten. Gleichzeitig



Vereinfachte Darstellung eines Reinforcement Learnings (copyright: public domain, CC0)

müssen bestimmte Annahmen etwa über bestimmte Charakteristika der Umwelt *a priori* implementiert werden, weil das POMDP sonst zu rechenintensiv wird und die Recheneinheiten (CPU/GPU) an Grenzen kommen. Damit ist auch ein Zielkonflikt definiert:

> *„Yet when one approximates, in any sense, there is no guarantee that a system will act in a precisely optimal manner. However we manipulate the mathematics, we pay a cost in one domain or another: either computationally or for the pursuit of the best outcomes.“*[1]

Die Herausforderung liegt also darin, ML-Modelle zu entwerfen, die *computationally frugal* sind und dennoch zu Ergebnissen kommen, die jedenfalls gut genug sind.

Zusammenfassend lässt sich festhalten, dass autonome Fahrzeuge auf der Basis von mathematischen Modellen Entscheidungen treffen, die *wahrscheinlichkeitsbasiert* sind. Entscheidungen können nicht auf der Basis von vollständiger Information über die Umwelt getroffen werden, sondern müssen mit Unsicherheit und (eigenen) Grenzen umgehen. Die Umwelt, in der sich die Maschine bewegen muss, ist dynamisch, d.h. sie verändert sich ständig und entsprechend muss auch die Maschine ihr Verhalten ändern. Sie kann nun insofern für diese Situationen „trainiert“ werden, als sie unter Anleitung, d.h. etwa durch Implementation eines Feedbacksystems, das Belohnungssignale für „gute“ Entscheidungen sendet, über die besten Entscheidungsfolgen in mannigfaltigen Szenarien informiert wird. Die Extrapolation aus diesem „Erfahrungsschatz“, der sowohl in der Computersimulation als auch im „real-world-testing“ gewonnen wird, verspricht, in zukünftige Situationen die „richtigen“ Entscheidungen zu treffen. Dabei ist gleichwohl nicht auszuschließen, dass die Maschine in Situationen gerät, zu denen sie keine Entsprechung in ihrem vorangegangenem „Wissen“ findet, so dass sie ggf. in einer nicht vorhergesehenen Art und Weise auf diese neue einzigartige Situation reagiert. Wenn sich AV-

Unternehmen also mit der Zahl ihrer gefahrenen Testkilometer brüsten, ist damit auch immer das Versprechen verbunden, dass die Maschine schon *alles gesehen hat* – unvorhergesehene Situationen, in denen die Maschine verwirrt ist und erratisch agiert, werden immer seltener und sollen schließlich gänzlich ausgemerzt sein: Fährt die Maschine immer und überall sicher, kann sie auf menschliche Hilfe verzichten und der höchste Autonomiegrad ist erreicht.

Ethik in einem größeren Zusammenhang: value functions

Man wird autonome Fahrzeuge nicht vernünftig und konsistent für Trolley-Probleme programmieren können: Weder muss die Maschine nur an *einem* Punkt entscheiden noch kann sie dies auf der Basis perfekten Wissens über die Konsequenzen der Entscheidung tun. Ferner können wir die Maschine auch gar nicht derart trainieren, dass wir ihr „schlechte" Beispiele und „gute" Beispiele für Dilemma-Lösungen zeigen und sie anhand dieser Daten lernt. Denn erstens gibt es per definitionem keine „Lösung" des Dilemmas, die nicht mit hohen Kosten verbunden wäre und zweitens gibt es ja überhaupt keinen Konsens darüber, was gute und schlechte Entscheidungen in Dilemma-Situationen sind – es gibt also überhaupt keinen passenden Datensatz, mit dem man die Maschine füttern könnte. Wenn man davon ausgeht, dass mit dem Advent des autonomen Fahrzeugs der Verkehr ohnehin sicherer wird, gerät die Beschäftigung mit Trolley-Problemen ohnehin in Begründungsschwierigkeiten: Warum sollte ein Phänomen, das heute bereits sehr selten ist, solche Prominenz zugesprochen werden, wenn es zukünftig *noch seltener* wird?

Wenn Trolley-Problem also zwar dazu taugen, menschliche Intuition für ethische Fragestellungen anzuregen, aber nicht dazu, autonome Fahrzeuge zu programmieren – welche ethischen Herausforderungen sollten wir uns stattdessen zuwenden?

Roff verweist auf die den Maschinen einprogrammierte Nutzenfunktion, der eine zentrale Bedeutung zukomme. Was der Nutzen sein soll, darüber muss es eine breite Diskussion geben:

> *„[W]e need to identify the values that we want to actualize through the engineering, design and deployment of technologies, like self-driving cars. There is thus a double entendre at work here: we know that the software running these systems will be trying to maximize their value functions, but we also need to ensure that they are maximizing society's too."*[2]

Es muss darüber gesprochen werden, welche Ziele die Maschine verfolgen soll und was passiert, wenn verschiedene Ziele in Konflikt zueinander geraten: Welche Ziele sind uns wichtiger? Denn das autonome

Fahrzeug sollte zwar einerseits versuchen, möglichst schnell von A nach B zu kommen, dabei aber andererseits möglichst keine Leitplanken touchieren.

Vor dem Hintergrund des Wissens darüber, wie autonome Fahrzeug *wirklich entscheiden*, liegt die Herausforderungen ferner darin, den Umgang mit Wahrscheinlichkeiten und die Gewichtung von Schadenszurechnungen zu bewerten: So käme der Vermeidung von schweren Verletzungen eine höhere Zielpriorität zu als der von leichten Verletzungen. Was aber, wenn die schwere Verletzung nur zu 30% wahrscheinlich ist, die leichte aber zu 70%? Derartige Gedankenspiele nähern sich gleichwohl wieder der Perspektive des Trolley-Problems an, suggerieren sie doch, dass die Maschine *einigermaßen sicher* einschätzen kann, wie Verletzungen ausfallen könnten. Wenn aber der Verkehr sicherer wird, die Zahl der Toten und Verletzen mit der Nutzung autonomer Fahrzeuge rapide abnimmt, sollte es dann nicht auch andere ethische Erwägungen zum autonomen Fahren jenseits der Zurechnung von etwaigen Schäden geben?

> *„From a value-sensitive design (VSD) standpoint, one may consider not only the question of lethal harm to passengers or bystanders, but a myriad of values like privacy, security, trust, civil and political rights, emotional well-being, environmental sustainability, beauty, social capital, fairness, and democratic value.“*[3]

So hat die Unternehmensberatung Deloitte in einer Studie aufgezeigt, wie autonome Transportfahrzeuge den Einzelhandelsmarkt in den USA revolutionieren könnten: Kunden sind nicht mehr darauf angewiesen, den nächsten Walmart anzusteuern, sondern werden von Zustellungsfahrzeugen angefahren und können an der eigenen Haustür ihre Einkäufe in Empfang nehmen. Menschen ohne eigenes Auto müssten nicht mehr auf andere Verkehrsmittel umsteigen, was in den USA oftmals eine echte Beeinträchtigung der Mobilität bedeutet. Damit könnten sich auch die Supermärkte auf dem Land verändern, wenn kaum noch „Fußkundschaft“ vorbei schaut – eine Transformation zu nicht-öffentlichen Logistikzentren a la Amazon wäre angedeutet. Gleichzeitig bestehen aber bereits heute in den USA sogenannte „food deserts“ – Regionen, in denen der nächste Supermarkt sehr weit entfernt nur mit dem Auto zu erreichen ist. Mit der AV-Technologie wäre es aufgrund der großen Distanzen nicht sonderlich profitabel diese entlgegenden Gegegenden anzusteuern. Sie werden gleichwohl insbesondere von wenig wohlhabenden Menschen bewohnt. Die Technik könnte hier also zu einer Verschlechterung ihrer Situation beitragen.

Diese mittelbaren Effekte einer Technologie, die verschiedenste Lebensbereiche zu durchdringen verspricht, haben immer auch ethische Dimensionen – etwa, wenn es um den gerechten Zugang zu Mobilitätsoptionen geht. Diese Effekte jenseits der konkreten und hoffentlich zukünftig als „Outlier“ zu verstehenden Unfallsituation

mitzudenken und einer ethischen Bewertung zu unterziehen, ist unerlässlich, wenn es darum geht, die Akzeptanz der Technik zu gewährleisten.

Reihe „Ethik und automones Fahren“

1. Was autonomes Fahren mit Ethik zu tun hat
2. Ethik & autonomes Fahren II: Trolley-Problem
3. Ethik & autonomes Fahren III: Das Problem mit dem Trolley-Problem
4. Ethik & autonomes Fahren IV: Wie autonome Fahrzeuge wirklich entscheiden
5. Ethik & autonomes Fahren V: Warum das Trolley-Problem doch wichtig ist
6. Ethik & autonomes Fahren VI: Ein selbstbestimmtes Ethik-Setting für mehr Akzeptanz?
7. Ethik & autonomes Fahren VII: Die deontologische Kritik an der Schadensoptimierung
8. Ethik & autonomes Fahren VIII: Recht und Dilemma
9. Ethik & autonomes Fahren IX: Die Ergebnisse der Ethik-Kommission „Automatisiertes und vernetztes Fahren“
10. Ethik & autonomes Fahren X: Ethik und Datenschutz

1. Roff, Heather M. 2018. The folly of trolleys: Ethical challenges and autonomous vehicles. Brookings Institution. ↵
2. ebd. ↵
3. ebd. ↵

← zurück weiter →

Über

Auf dem Autonoblog schreiben Wissenschaftler unterschiedlicher Disziplinen über sozialwissenschaftliche, ethische wie rechtliche Aspekte des autonomen Fahrens. Unter Leitung von Dr. David Knollmann und Prof. Volker Lüdemann informiert das Autoren-Team regelmäßig über neueste Entwicklungen rund um das autonome Fahren. Der Autonoblog ist eine Publikation des Niedersächsischen Datenschutzzentrums (NDZ), einer wissenschaftlichen Einrichtung der Hochschule Osnabrück, und des vom Europäischen Fonds für regionale Entwicklung geförderten Forschungsprojekts „Demokratie des Fahrens – Sollen Autos moralische Entscheidungen treffen?“ (DeFrAmE).

Suche …

Tags

AI Apple Argo AI Arizona Aurora Automatisierungsstufen Autopilot BMW CES2019 Daimler DB Schenker Dilemma Distronic Plus DMS Einride Elon Musk Ethics Ethik Ford General Motors Geotonomy Hype Lidar Lyft machine learning Mercedes Benz ML Model S NHTSA NTSB SAE SAE-Levels Sicherheit Steve Wozniak SuperCruise Taxi-Service Tesla Trolley-Problem Trolley-Probleme Trolley-Problems Uber Volkswagen Waymo Wiki Zukunftsprognose

Kategorien

Analyse

Neueste Beiträge

Archive