

# Autonoma bilar: AI och etik

Philip Lindfors

Kandidatavhandling i datateknik

Handledare: Marina Waldén

Fakulteten för naturvetenskaper och teknik

Åbo Akademi

Våren 2020

Referat

TODO

# Innehållsförteckning

1	Inledning.....	1
2	Nivåer av automation .....	2
2.1	Icke-autonoma nivåer .....	2
2.2	Autonoma nivåer .....	3
3	AI och maskininlärningsmetoder för autonoma bilar.....	3
3.1	Faltningsnätverk (CNN).....	5
3.2	Återkommande neuralt nätverk (RNN).....	6
3.3	Djupförstärkt maskininläring (DRL).....	7
3.4	End2End-learning.....	8
4	Etiska aspekter.....	9
4.1	Etikens betydelse.....	9
4.2	Spårvagnsproblemet .....	11
4.3	Kollisionsundvikning .....	11
5	RUBRIK.....	13
5.1	Kraschoptimeringsalgoritm .....	13
5.2	Den etiska beslutsprocessen .....	15
6	Diskussion och sammanfattning.....	16
	Litteraturförteckning .....	16

# 1 Inledning

Världen är beroende av fordon. Människor kör bil till jobbet, barn tar bussen till skolan, varor transporteras med långtradare. Hela samhället förlitar sig på existensen av dessa fordon och även om så gott som alla bilar i trafik ännu behöver en människa bakom ratten, är idén om autonoma bilar gammal.

En autonom, alternativt självstyrande, bil är per definition en bil ”navigerad och manövrerad av en dator utan behov av mänsklig kontroll eller ingripande under en rad körsituationer och förhållanden” [1]. Det första steget mot autonoma bilar skedde redan år 1925, då radioutrustningsföretaget Houdina Radio Control satte fast en antenn på bakkdelen hos en bil, som sedan tog emot radiovågor från en bakomkörande bil. Antennen skickade sedan signaler till ett par strömbrytare som hanterade små elektriska motorer som styrde bilens rörelser [2]. Tekniskt sett kan detta inte betraktas som en självstyrande bil, utan snarare en bil som styrdes av en förare utanför bilen. Det påvisar ändå det faktum att självstyrande bilar som koncept inte är något nytt.

Sedan dess har teknologin kommit en lång väg. Den en gång i tiden avlägsna drömmen om en självstyrande bil är inte längre än dröm. Det som vi undrar idag är inte om, utan när en självstyrande bil kommer att vara den nya standarden i trafiken. I takt med att automatiseringen utvecklas och gradvis tar över vår vardag, försvinner också den mänskliga faktorn steg för steg. Detta gäller också bilar: ifall framtida fordon är fullständigt autonoma och aldrig kräver ett enda mänskligt ingripande under körningen, betyder det också att fordonet måste kunna bedöma varje trafiksituation och därefter reagera på situationen åtminstone lika bra som en människa. I trafiken uppstår det dock ibland situationer som inte är så svartvita, och kräver mer av föraren än att hen bara följer trafikmärken och lagar. Föraren kan hamna i en situation där det inte finns något självklart svar på hur denne ska reagera, utan är tvungen att snabbt ta ett moraliskt beslut.

Hur hanterar en helt självstyrande bil dessa situationer? Kan en bil programmeras till att inte bara följa trafikmärken och -lagar, utan också tillämpa de etiska principerna som människor intuitivt följer? Är det ens möjligt att i kontrollalgoritmer hos autonoma bilar återskapa hela det mänskliga beteendet?

Jag kommer i denna avhandling först att ge en överblick av de AI-metoder som idag används för att programmera bilar, och sedan närmare försöka besvara frågorna ovan.

## 2 Nivåer av automation

Society of Automotive Engineers, förkortat SAE, publicerade år 2014 en standard vid namn SAE J3016 där man definierar sex olika nivåer av automation, från 0 (ingen automation) till 5 (fullständig automation) [3]. Vidare kan dessa sex nivåer delas in i två delar: icke-autonoma samt autonoma nivåer.

### 2.1 Icke-autonoma nivåer

Vid de icke-autonoma nivåerna 0, 1 och 2 så har den mänskliga föraren full kontroll över fordonet. Även om en viss assistans kan finnas, såsom farthållare eller spårkontroll, så ligger det slutgiltiga ansvaret alltid hos föraren. Föraren måste konstant vara uppmärksam och övervaka de hjälpmedlen som finns i bilen. Nivå 0 innebär att alla funktioner hos bilen handskas av föraren, som manuellt måste kontrollera hastigheten och bromsarna [3]. Även om bilen har en grundläggande farthållare så räknas det som nivå 0, eftersom farthållaren endast är aktiv om den aktiveras och ställs in av en mänsklig hand.

Vid nivå 1 finns vissa assisterande hjälpmedel så som adaptiv farthållning eller parkeringssensorer. Till skillnad från en enkel farthållare, så kan en adaptiv farthållare självständigt justera bilens hastighet. Enkla parkeringssensorer varnar föraren ifall bilen detekterar ett objekt i bilens riktning, men föraren måste till fullo manövrera bilen vid parkeringen. Ifall spårkontroll hör till bilens funktioner innebär det också en automation av nivå 1, så länge bilen inte har en adaptiv farthållare. Om bilens funktioner innehar både spårkontroll och adaptiv farthållning, räknas det som en automationsgrad av nivå 2. Denna nivå kallas också partiell automation, där två eller flera funktioner används samtidigt för att assistera körningen av fordonet [3]. I dagens läge är nivå 2 den högsta graden av automation hos nya bilar.

## 2.2 Autonoma nivåer

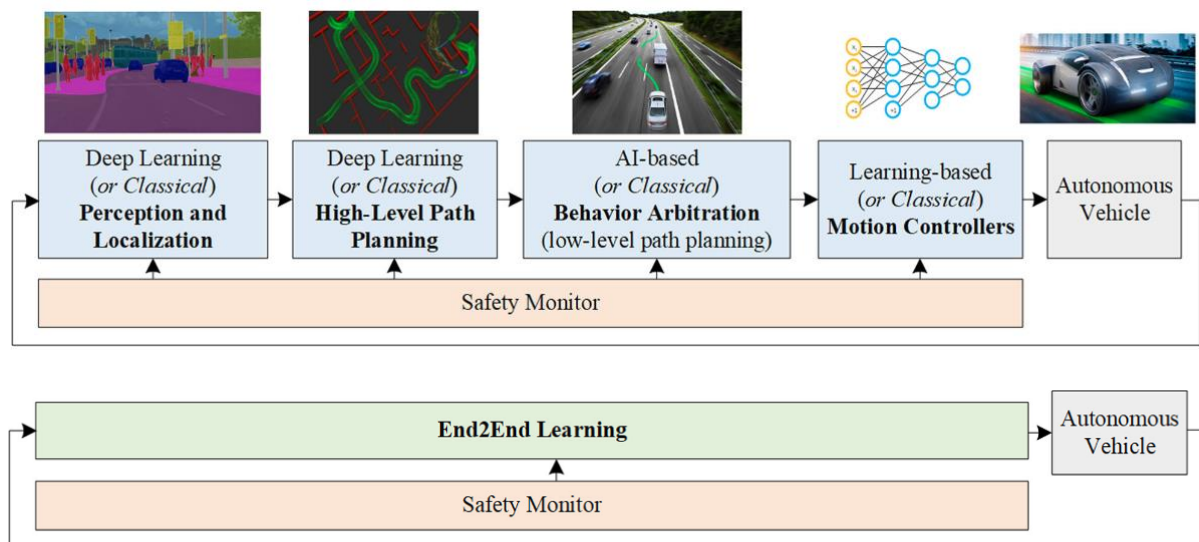
De autonoma nivåerna 3, 4 och 5 innebär att föraren inte själv kör bilen när de automatiska körfunktionerna är aktiverade, även om föraren sitter bakom ratten. Vid nivå 3 har bilen förmågan att genom sensorer ta beslut utan mänskligt ingripande. Föraren måste ändå vara beredd att ta över kontrollen av fordonet ifall något oförutsett skulle hända. Ett fordon på nivå 3 kan bara ha kontroll över körningen under vissa situationer och när ett par specifika villkor uppfylls, medan ansvaret flyttas över till föraren vid alla andra situationer [3].

Bilar på nivå 4, också kallad hög automation, kan under de flesta förhållandena vara självstyrande utan mänskligt ingripande. Det finns dock precis som i nivå 3 situationer där fordonet inte klarar av körningen utan kräver att föraren måste ta över kontrollen, till exempel när väglaget eller vädret inte är optimalt.

Nivå 5, eller full automation, innebär att bilen är fullständigt självstyrande i alla situationer och förhållanden. Människan i bilen behöver inte under några omständigheter ta över kontrollen av bilen [3]. Eftersom bilen inte längre behöver en mänsklig förare, elimineras också behovet av utrustning såsom ratt och pedaler.

## 3 AI och maskininlärningsmetoder för autonoma bilar

I det här avsnittet kommer jag att behandla hur AI och maskininläring används i autonoma bilar. Autonoma bilar bearbetar en mängd observationer och information som fås från flera apparaturer tillhörande bilen: kameror, radars, LiDAR (*light detection and ranging*), positionssystem samt olika typer av sensorer. Den här informationen används sedan av bilen för att göra beslut hur den ska köra. Besluten kan kalkyleras på två sätt. Det ena är att beräkningarna görs i en modulär pipeline (alternativt 'pipa') och den andra möjligheten är att man använder sig av End2End-learning [4]. I Figur 1 finns en illustration av de två olika tillvägagångssätten.



Figur 1. Modulär pipeline och End2End-inläring [4]

Modulär design innebär en pipeline som är hierarkiskt indelad i fyra komponenter:

1. **Uppfattning och lokalisering** (*perception and localization*). Den här komponenten är en fundamental funktion för en autonom bil. Den ger bilen information om köromgivningen (*driving environment*) som t.ex. var den kan köra, positionen för omgivande objekt, andra bilars hastigheter och även en prediktion om framtida tillstånd.
2. **Ruttplanering på högnivå** (*high-level path planning*). Bilen ska ha förmågan att hitta en optimal och säker rutt mellan startpositionen A och slutdestinationen B. Bilen måste ta hänsyn till alla hinder som finns i omgivningen och beräkna en kollisionsfri rutt till slutdestinationen.
3. **Beteende och ruttplanering på lågnivå** (*behavior arbitration, low-level path planning*). I samband med att bilen tar sig fram på den planerade ruten, så måste den följa trafiklagarna och interagera med andra trafikanter på ett säkert sätt.
4. **Rörelsekontroller** (*motion controllers*). Dessa kontroller omvandlar bilens avsikt till en konkret handling. Det huvudsakliga målet är att exekvera den planerade avsikten genom att förse nödvändig input till hårdvaran som sedan genomför handlingen.

I den här typen av design görs beräkningarna antingen genom användning av artificiell intelligens (AI) och maskininläring, eller genom klassiskametoder som inte är

sjävlärande. End2End-inläring är huvudsakligen baserat på djupinläring och innebär att sensorisk information överförs (*mapping*) direkt till utstyrsignalerna [4].

Till näst kommer jag att ge en kort överblick av de tre vanligaste djupinlärningsstrategierna som används i självstyrande bilar [4]: faltningsnätverk (*convolutional neural network, CNN*), återkommande neurala nätverk (*recurrent neural networks, RNN*) samt djupförstärkt maskininläring (*deep reinforcement learning, DRL*). Till sist kommer jag också att beskriva End2End-learning.

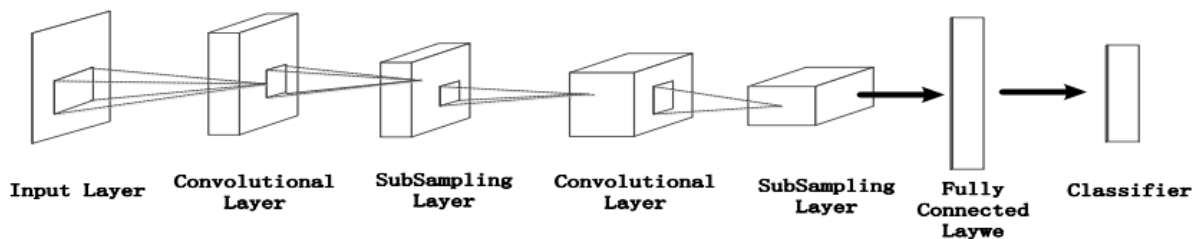
### 3.1 Faltningsnätverk (CNN)

Faltningsnätverk är en typ av djupt neuralt nätverk (*deep neural network, DNN*) som ofta används i objekt-detektering. Faltningsnätverk består vanligtvis av fyra olika typer av skikt [5] (se Figur 2):

1. Inputlagret, som tar in en inputvolym i tre dimensioner (en bild).
2. Faltningslagret som använder sig av ett flertal filter för att extrahera olika kännetecken från inputbilden.
3. Sammanslagningsslagret (*pooling layer*) som reducerar den rumsliga storleken (*spatial size*) av framställningen för att reducera antalet parametrar.
4. Det till fullo anslutna lagret som sammanbinder alla neuroner till alla aktiveringar i det tidigare skiktet.

I autonoma bilar så används faltningsnätverk för att bearbeta rumslig information som till exempel bilder. När ett objekt väl är detekterat, kan spårningsteknologi som kameror och LiDAR användas för att spåra närliggande fordon eller fotgängare för att säkerställa att fordonet inte kolliderar med rörliga objekt [6]. Fördelen med att använda faltningsnätverk framom andra klassificeringstekniker eller neurala nätverk är att nätverket effektivt kan tolka mönster och strukturell information i en bild [7]. En annan fördel är att faltningsnätverket kan utnyttja 2-D-strukturen hos bilder och har således en högre precision än ett standard neuralt nätverk. Faltningsnätverk är idag standard för objektidentifiering, men används också inom andra områden som till exempel naturlig språkbehandling, ljudanalys och textklassificering. I figur 2 illustreras ett typiskt faltningsnätverk.

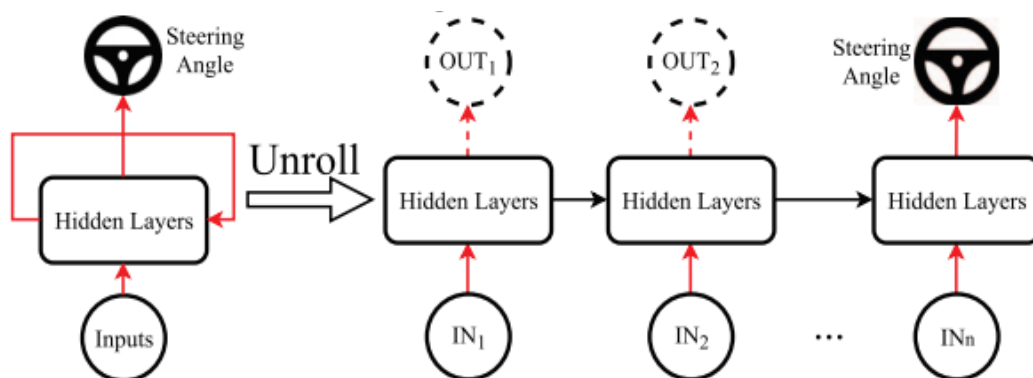




Figur 2. Typisk representation av ett faltningsnätverk [8](egen översättning)

### 3.2 Återkommande neuralt nätverk (RNN)

Ett återkommande neuralt nätverk (*recurrent neural network, RNN*) är en typ av artificiellt neuralt nätverk som är speciellt bra på att behandla temporär sekventiell data, såsom text eller videoflöde [4]. Till skillnad från CNN så tillåter RNN slingor i nätverket; varje lagers output matas inte bara till det följande lagret, utan också till det föregående lagret. Ett sådant arrangemang tillåter prediktionsoutput för tidigare inputs, som till exempel tidigare bildrutor i en videosekvens, som också ska beaktas vid prediktion av den aktuella inputen [9]. Den här processen upprepas tills en slutgiltig output är erhållen. Figur 3 visar en förenklad arkitektur hos återkommande neuralt nätverk. Den utvecklade versionen demonstrerar hur slingan tillåter en sekvens av inputs (bilder) bli matad till nätverket och styrvinkeln förutspås då baserat på bilderna.



Figur 3. En förenklad RNN-arkitektur. [9]

Risken med användningen av slingor i nätverket är att det kan leda till en felaktig modell. LSTM (*long-short-time-memory*) är en typ av RNN som är designat för att lösa det här problemet [9]. LSTM kan användas t.ex. för att förutse komplexa fordonsrörelser i flerfiliga vägförbindningar [10]. LSTM kan också kombineras med andra typer av nätverk.

En modell som kallas Chauffeur model, består av en CNN-modell för att extrahera kännetecken och drag från en bild, kombinerat med en LSTM-modell för att förutse fordonets styrningsvinkel. CNN-modellens input är en bild, medan LSTM-modellens input består av en sammanbindning av 100 kännetecken (*features*) extraherade av CNN-modellen från de tidigare 100 konsekutiva bilderna [9].

### 3.3 Djupförstärkt maskininlärning (DRL)

Djupförstärkt maskininlärning (*deep reinforcement learning*) är en kombination av djupinlärning och förstärkningsinlärning som används för att skapa effektiva algoritmer som kan tillämpas inom områden såsom AI, finans och hälsovård [11]. Kort sammanfattat kan förstärkt maskininlärning beskrivas som att lära sig genom försök och misstag (*trial and error*).

I ren förstärkningsinlärning finns ingen extern feedback, utan agenten (i detta fall en självstyrande bil) måste självmant besluta hur den ska agera i en godtycklig situation. För att detta ska lyckas måste bilen lära sig av erfarenhet. Det måste dock finnas något slags belöningsystem så att bilen vet om beslutet den just tog var positivt eller negativt. Förstärkningsinlärning använder sig av ett system där det rätta beslutet belönas, medan det felaktiga beslutet bestraffas. Således förstärks bilens beteende. Orsaken till att deep reinforcement learning används i autonoma bilar är att bilen kontinuerligt skall kunna lära sig av varje erfarenhet genom att få feedback på varje beslut den gör, och därmed ha den kunskap som krävs för att kunna ta rätt beslut i framtida situationer [7].

Inom DRL finns en variant som kallas Deep Q-learning. I vanlig Q-learning kan agenten vara i ett antal lägen och utföra ett antal handlingar. Vid vilken tidpunkt som helst, befinner sig agenten i ett specifikt läge. I nästa iteration, kan agenten byta läge genom att utföra en handling. Den här handlingen följs sedan av endera belöning eller bestraffning och agentens mål är att maximera belöningarna [12]. Begreppet Q-learning härstammar från det faktum att namnet på den funktion som returnerar belöningen är Q. Eftersom Q-värdet inte är känt så används en icke-linjär funktionsapproximation som t.ex. ett neuralt nätverk för att estimerar Q-funktionen. Det neurala nätverket som används för att uppskatta Q-funktionen kallas för ett Deep Q-nätverk (DQN). Det är den här

kombinationen av förstärkningsinlärning (Q-learning) och djupinlärning (neurala nätverk) som skapar en djup förstärkningsinlärning [13].

### 3.4 End2End-learning

End2End-inlärning, alternativt end-to-end-inlärning, innebär att sensorisk data mappas direkt till styrkommandot. Inputen tas från bilder och punktmoln (*point clouds*). Punktmoln är datamängder som representerar objekt eller rum. Punkterna representerar X-, Y-, och Z-koordinaterna för en enskild punkt av en underliggande yta. Genom detta samlas ett stort antal enskilda spatiella mått in och slås samman till en enhetlig datamängd som då representerar hela ”rummet”. Inputen från bilderna och punktmolnen mappas sedan direkt till styrkommandot. Det här kan jämföras med den modulära pipelinen där objekt först detekteras, sedan planeras en rutt, och till sist beräknas rörelsekontrollen [4].

Användningen av faltningsnätverk är speciellt effektiv vid en tillämpning av End2End-inlärning. Bojarski et al [14] utvecklade en End2End-metod där ett faltningsnätverk lärde sig att överföra råa pixlar från en enskild, framåtriktad kamera direkt till styrkontrollerna. Det här visade sig vara mycket effektivt: trots minimal träningsdata från människor så lärde sig bilen att köra i trafik, både på småvägar utan filmarkeringar men också på stora motorvägar. Den klarade också av att navigera i områden såsom parkeringsplatser eller vägar som inte var asfalterade. Allt detta genomfördes med endast den mänskliga styrsignalen som träningsdata, där en människa manuellt körde bilen i mindre än 100 timmar på olika typer av vägar och under varierande väderförhållanden. Faltningsnätverket var alltså kapabelt till att lära sig att följa vägar och hållas i den rätta filen utan att dela upp inlärningsprocessen i mindre bitar som objektidentifiering, ruttplanering och rörelsekontroller.

## 4 Etiska aspekter

I det här avsnittet kommer jag att behandla de etiska aspekterna som uppstår i samband med autonoma bilar. Mänskliga förare lär sig kontinuerligt från till exempel sina misstag och kan baserat på sin erfarenhet snabbt tolka ovanliga trafiksituationer baserat på sammanhanget. En självstyrande bil kan ha problem med att förstå situationer den aldrig tidigare blivit utsatt för, och dess reaktion till situationen kan vara väldigt olik jämfört med hur en människa skulle ha reagerat.

### 4.1 Etikens betydelse

Ifall framtida fordon kommer att vara fullständigt autonoma innebär det också att de kommer att vara tvungna att i varje situation kunna återskapa det mänskliga beteendet. Dock uppstår det ibland farosituationer där det mänskliga beslutet inte tas som en följd av logiskt tänkande, utan istället är det en form av ett panikartat agerande som kan betraktas som en impuls snarare än ett aktivt beslut.

Patrick Lin, en forskare som närmare undersökt hur etik och teknologi är sammanlänkade, ger ett exempel [15]: ditt självstyrande fordon råkar ut för ett mardrömsscenario där en flicka i lågstadieåldern står mitt på vägen tillsammans med hennes farmor. Ifall fordonet svänger till vänster kör det över flickan, medan en högersväng resulterar i att farmorn blir överkörd. Ifall fordonet inte svänger alls, kör det över båda två. Det mänskliga beslutet kan inte kalkyleras, eftersom beslutet skulle tas på bråkdelen av en sekund utan betänketid. Därmed är det svårt att ställa en människa inför rätta och beskylla hen för att ha agerat felaktigt, eftersom allting hände så snabbt.

Den autonoma bilen kan dock i utvecklingsfasen programmeras till att alltid agera på samma sätt ifall en likande situation som ovan skulle inträffa. Problemet är att det finns inget entydigt svar på vad som vore det korrekta beslutet etiskt sett. Många anser det som en självklarhet att den unga flickan har större rätt till att överleva eftersom hon har hela sitt liv framför sig, medan farmorn redan har haft ett långt liv fyllt av erfarenheter. Detta går dock emot IEEE:s löfte om att ”behandla alla personer likvärdigt och inte diskriminera någon på grund av dennes ras, religion, kön, funktionshinder, ålder, nationalitet, sexuella läggning eller könsidentitet” (direkt översättning). Ifall bilen är programmerad till att

välja offer endast på basis av ålder, är detta ren och skär åldersdiskriminering. Ifall man tolererade åldersdiskriminering, vad finns det som säger att andra former av diskriminering inte är tillåtna?

Ett alternativ i det här dilemmat vore att inget beslut alls tas, utan bilen fortsätter helt enkelt sin färdriktning även om det skulle innebära att båda personerna dör. Etiskt sett vore det här den enklaste lösningen eftersom inget aktivt val görs över vem som har rätten att överleva, men samtidigt känns det självklart att två dödsfall är värre än ett, även om ”fel” person skulle omkomma. Ett annat alternativ vore att bilen helt enkelt slumpmässigt svänger till höger eller vänster utan att beakta några yttre faktorer. Därmed görs aldrig ett aktivt beslut, utan den som överlever hade helt enkelt turen på sin sida. Detta verkar dock heller inte som någon bra lösning, eftersom slumpen avgör vem som överlever även om det är möjligt att det fanns många orsaker till varför det vore bättre att göra ett visst val framom ett annat.

Situationer som den ovan kan kännas osannolika och man kan tänka sig att de aldrig kommer att uppstå på riktigt, men samtidigt påvisar de också det faktum att autonoma bilar borde innehålla åtminstone någon form av etiska regler att utgå från. Samtidigt är det i princip omöjligt att säga vilken form av etisk skola den autonoma bilen ska följa, när inte heller människor kan enas om vad som är det etiskt korrekta beslutet. I en undersökning [16] ställdes ett antal människor frågan om hur de skulle agera i en situation där den autonoma bilen hade två val: antingen köra över tio fotgängare men rädda fordonets passagerare, eller så köra in i en vägg med utgången att passageraren dör medan de tio fotgängarna överlever. Undersökningen visade att 76% av de tillfrågade ansåg att det rätta valet vore att offra passageraren, medan 24% ansåg att fordonet alltid bör prioritera den egna passagerarens liv.

Det ironiska i undersökningen är att även om majoriteten svarade enligt den utilitaristiska etiska skolan (maximera lycka, minimera lidande), så ansåg majoriteten att de inte skulle köpa en bil ifall den offrade passagerarna framom ett större antal fotgängare. De tillfrågade ansåg alltså att utilitaristiska fordon var de mest etiska, men att de samtidigt själva hellre skulle köpa en bil som alltid prioriterade passagerarnas liv. Detta är en aspekt som biltillverkarna också måste reflektera över. Ifall bilarna skulle programmeras till att alltid minimera antalet offer oberoende av vem som dör, skulle det också antagligen resultera i sämre försäljningssiffror i jämförelse med om man programmerade bilarna till att alltid prioritera passagerarnas liv.

## 4.2 Spårvagnsproblemet

Ett scenario som ofta tas upp är ett gammalt tankeexperiment kallat spårvagnsproblemet (eng. the trolley problem). Kort förklarat så består situationen av en spårvagn som okontrollerat skenar iväg på ett spårvagnsspår och är på väg att köra över fem intet ont anande människor som står längre fram på banan. Självt står du vid en spak som byter vagnens spår: om du drar i spaken, byter vagnen spår och du räddar fem liv. På det andra spåret står dock en person som skulle bli överkörd ifall du beslöt dig för att dra i spaken. Det etiska dilemma är här att ifall du väljer att dra i spaken, har du aktivt dödat en människa, medan om du lät vagnen fortsätta på samma spår så har du endast låtit människor dö, om dock i en större mängd.

Patrick Lin applicerade detta dilemma på en autonom bil [15]: ponera att du manuellt kör en bil som dock har förmågan att köra sig själv. En bit framför dig i samma riktning har du fem fotgängare som du är på väg att köra över på grund av till exempel ouppmärksamhet. Det här upptäcker det autonoma systemet hos bilen, och för att undvika kollision tar det då över kontrollen och svänger snabbt till höger vilket resulterar i en (1) människas död. Ifall systemet låter föraren ha kontrollen dör fem människor, men bilen har här inte aktivt dödat någon utan endast låtit människor dö genom passivitet. Ifall systemet tar kontrollen och svänger till höger har bilen aktivt dödat en människa.

Hur ska bilens beteende vara programmerat att hantera en sådan här situation? Det kan argumenteras för att det objektivt är bättre att skada minsta möjliga antal människor. Bilen skulle i såfall många gånger göra ett aktivt val att döda en människa eftersom den strävar efter att objektivt sett skapa minst skada. Ifall bilen vore programmerad till att alltid skada minsta möjliga antal människor så skulle det kunna uppstå situationer där bilen bedömer att det bästa utfallet sker ifall den kör in i en vägg, med det möjliga resultatet att bilens passagerare dör. Detta skulle knappast fungera i praktiken eftersom ingen skulle vilja köpa en bil som innehar den här risken.

## 4.3 Kollisionsundvikning

Ifall en självstyrande bil i en ovanlig situation kunde undvika en kollision genom att t.ex. bromsa eller svänga åt sidan så borde den rimligtvis göra det, eftersom bilens passagerare då kommer ur situationen utan skada. Det finns dock även här situationer som inte är så svartvita. Ponera att bilen står stilla vid ett trafikljus och väntar på att fem barn skall gå över vägen. Bilen upptäcker att det bakifrån kommer en annan bil i hög fart som är på väg att kollidera med dig. En kollision bakifrån skulle högst troligen inte resultera i dödlig utgång för passageraren även om personskador nog kunde uppstå. Ifall bilen var programmerad till att alltid undvika kollisioner ifall detta kan göras på ett säkert sätt, så skulle den snabbt svänga åt något håll för att undvika att bli påkörd bakifrån. I teorin kan detta verka som att bilen är programmerad till att göra det rätta beslutet, eftersom bilen helt och hållet undviker skada och tar passageraren skadefri ur en farlig situation. I den här situationen skulle dock bilens beslut leda till att de fem barnen som går över vägen blir påkörda av den andra bilen. Därmed har bilen varit programmerad till att prioritera flera möjliga dödsfall framför endast person- och materiella skador, sålänge det innebär att passageraren kommer helt skadefri ur situationen.

Det andra utfallet är att den självstyrande bilen inser att ifall den svänger undan, kommer barnen att dö eller åtminstone skadas allvarligt. Om den istället står kvar och inväntar kollisionen kommer det att uppstå materiella skador och möjliga personskador för bilens passagerare, men risken att någon ska dö är mycket mindre än om bilen svängde ur vägen. Det är svårt att säga hur en mänsklig förare skulle ha reagerat i en sån här situation. En människa måste dock göra det här beslutet i stunden på samma gång det inträffar, medan en autonom bils beslut i princip görs redan i programmeringsfasen av mjukvaruutvecklarna. Därmed kan det argumenteras för att ett mänskligt beslut, om än ”fel”, är mer förlåtligt än om bilen från början vore programmerad till att till varje pris undvika en kollision ifall detta kan göras på ett (för bilen) säkert sätt.

Det finns en mängd andra scenarion som vidare påpekar det faktum att autonoma bilar ännu har en bit att gå innan de kan bli en del av vardagen. Samtidigt kan det argumenteras för att det inte finns något entydigt svar till dessa etiska dilemman [17]. Även om en mänsklig förare befann sig i ett etiskt dilemma kan man argumentera att vilket beslut denne än tar så kan det betraktas som ett etiskt felaktigt beslut, och hur ska man då kunna kräva att en autonom bil i samma situation alltid ska ta det rätta beslutet? Vidare så är dessa scenarion väldigt osannolika och kan betraktas närmast som irrelevanta jämfört

med ofta förekommande trafiksituationer, där en autonom bil i många fall kan vara säkrare än en bil kontrollerad av en människa.

## 5 RUBRIK

Även om de situationer som tagits upp i föregående kapitel inte kan betraktas som ofta förekommande, kommer man inte undan det faktum att med tillräckligt många autonoma bilar i trafiken så kommer olyckor att inträffa på ett sätt eller annat. Bilen kan bli tvungen att välja mellan två eller flera dåliga alternativ som alla har en dödlig utgång för någon av parterna. Forskning indikerar att självstyrande bilar kommer att vara inblandade i olyckor mycket mera sällan än bilar kontrollerade av människor, men trots det är olyckor oundvikliga och dess följder har etiska konsekvenser [18]. Därför har industrin i viss mån börjat beakta algoritmerna för den så kallade etiska beslutsprocessen.

Kontrollalgoritmerna som avgör hur en autonom bil agerar kan komma att grundligt granskas efter att algoritmernas funktion helt eller delvist orsakat en allvarlig olycka. Det här innebär att ett stort ansvar läggs på de personer som tar fram dessa algoritmer för att garantera att kontrollalgoritmerna fungerar på ett sådant sätt som för människor är acceptabelt ur ett lagenligt och etiskt perspektiv [19].

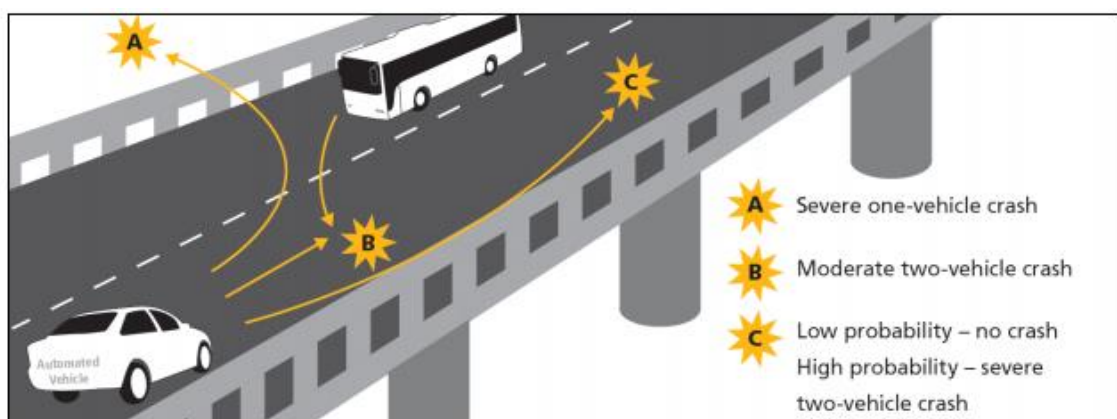
### 5.1 Kraschoptimeringsalgoritm

Eftersom olyckor är oundvikliga borde självstyrande bilar ha en algoritm som beräknar hur bilen ska agera vid en kollision: bilen måste vara programmerad att krascha på ett så säkert sätt som möjligt. Algoritmen för kraschoptimering innebär kortfattat att vid en oundviklig krasch försöker algoritmen optimera kraschen så att minsta möjliga skada uppstår till följd av den. Det kan argumenteras för att vid en sådan här situation kunde bilen ge över kontrollen till människan, men detta är i majoriteten av fallen en dålig lösning eftersom den mänskliga reaktionstiden är lång jämfört med AI. Dessutom är det svårt för en människa att snabbt byta fokus från en sak till en annan. Därför måste bilen vara beredd att agera så säkert som möjligt när en kollision är på väg att inträffa [20].



Det kan också dras paralleller mellan en kraschoptimeringsalgoritm och en inriktningssalgoritm (*targeting algorithm*). Vid en optimal kollision borde bilen rikta in sig på ett mål med utgången att minsta möjliga skada uppstår. Det här torde i praktiken innebära att bilen i de flesta fall skulle sträva efter att medvetet och systematiskt kollidera med t.ex. stora stadsjeepar istället för mindre bilar. De stora fordonens passagerare skulle därmed vara utsatta för kollisioner i högre grad än andra bilister endast av den anledningen att det objektivt sett är det bästa alternativet. En stor barnfamilj som behöver en stor och säker bil med mycket utrymme skulle alltså ironiskt nog ha större risk att vara inblandade i trafikolyckor än om de använde sig av en mindre bil.

Hur borde då bilen vara programmerad att agera i oundvikliga olyckor? I figur 4 illustreras en situation där en autonom bil kör på en tvåfilig bro då den noterar att den mötande bussen är på väg att svänga in i fel fil. Den autonoma bilen måste då ta ett beslut på basis av hur den är programmerad att hantera liknande situationer och har i det här fallet tre alternativ. Det första alternativet är att bilen svänger åt vänster och därmed kör av bron vilket skulle resultera i en garanterad kollision. Det andra alternativet är att bilen inte ändrar riktning utan kör rakt fram med utgången att bilen och bussen frontalkrockar. Det tredje alternativet är att bilen håller sig så långt till höger som möjligt och hoppas att bussen hinner ändra sin riktning och svänger tillbaka till sin egen fil före en kollision inträffar, något som i denna situation kan betraktas som osannolikt om än dock möjligt. Det är betydligt mera sannolikt att bussen fortsätter sin riktning till fel fil vilket då resulterar i en kollision.



Figur 4. Tre möjliga utfall då en buss plötsligt börjar köra över till motsatt fil [21]

I den här situationen finns det endast ett scenario (som dessutom är väldigt osannolikt) som resulterar i att den autonoma bilen och dess passagerare tar sig ut ur situationen oskadda. De övriga utfallen resulterar i materiella skador, personskador och i värsta fall dödsfall. Därför behövs någon typ av kraschoptimeringsalgoritm som för varje liknande situation beaktar alla faktorer och sannolikheten för varje alternativ och tar ett beslut utgående från den informationen.

## 5.2 Den etiska beslutsprocessen

Ifall även de mest utvecklade och sofistikerade autonoma fordonen kommer att vara inblandade i trafikolyckor, behöver det finnas något typ av system som behandlar den etiska beslutsprocessen d.v.s. vilket alternativ som ur ett etiskt perspektiv är det lämpligaste. Till skillnad från en mänsklig förare som tar ett beslut i samma ögonblick som en farlig situation uppenbarar sig, är det autonoma fordonets beslut taget månader eller till och med år på förhand. Fordonet tar ett beslut baserat på den information som i stunden är tillgänglig, men själva beslutet är ett resultat av hur fordonet är programmerat att agera i liknande situationer som den situation fordonet för tillfället befinner sig i.

Det finns ingen uppenbar lösning, men genom någon form av etisk bas för hur fordonet ska agera i trafiken skulle man ta ett steg närmare en lösning. Goodall föreslår tre faser som kunde implementeras i utvecklingsprocessen [21]. Den första fasen är *rationell etik*, vilket innebär att systemet skulle belöna beteende som minimerar omfattande skada. Det här innebär i praktiken att personskador är att föredra framom dödsfall, materiella skador framom personskador samt att människor i riskzonen skall skyddas (t.ex. fotgängare).

Den andra fasen, en blandning av *rationell etik* och *artificiell intelligens*, kräver mjukvara som i dagsläget inte är tillgänglig. I den här fasen så kan mjukvaran i en autonom bil använda sig av maskininlärning för att korrekt tolka etiska beslut samtidigt som den är bunden till att följa riktlinjerna i den första fasen.

Den tredje och sista fasen är *feedback genom naturligt språk*. En nackdel med neurala nätverk, som skulle användas i den andra fasen, är dess oförmåga att ge en klar bild av varför det agerade som det gjorde. Ett neuralt nätverk är svårt att demontera (*reverse engineering*) och det är därför komplicerat att fastställa hur nätverket kom fram till sitt beslut. Det är extremt viktigt att förstå logiken bakom det autonoma fordonets agerande

vid olyckor, speciellt om fordonet har agerat på ett oväntat sätt. Ifall detta inte kommer till kännedom, så är det omöjligt att garantera att samma problem inte uppstår igen. För att få mer kunskap om varför ett neuralt nätverk agerar som det gör, så har datavetare utvecklat teknik som extraherar information från det neurala nätverket och gör det mera lättförståeligt för människor. I praktiken så översätter den här processen nätverkets interna kunskap till en mängd symboliska regler, som sedan kan uttryckas i ett naturligt språk. Den här tekniken är dock inte perfekt, men fungerar som en bra bas för framtiden.

## 6 Diskussion och sammanfattning

TODO

### Litteraturförteckning

- [1] S. Yakamoto och H. Mori, *Human Interface and the Management of Information: Information in Applications and Services*, Tokyo: Springer International Publishing, 2018.
- [2] K. Bimraw, "Autonomous Cars: Past, Present and Future," Patiala.
- [3] S. International, "Levels of Driving Automation," 07 01 2019. [Online]. Available: <https://www.sae.org/news/2019/01/sae-updates-j3016-automated-driving-graphic>. [Använd 02 03 2020].
- [4] S. Grigorescu, B. Trasnea, T. Cocias och G. Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of Field Robotics*, 2019.
- [5] J. Patterson och A. Gibson, *Deep Learning: A Practitioner's Approach*, O'Reilly Media, 2017.
- [6] J. Tang, S. Liu, S. Pei, S. Zuckerman, C. Liu, W. Shi och J.-L. Gaudiot, "Teaching Autonomous Driving Using a Modular and Integrated Approach".
- [7] C. Syed Owais Ali, S. Riaz, M. Bilal Zaib och M. Nauman, "Self-driving Cars Using CNN and Q-learning," Peshawar, Pakistan, 2018.
- [8] S. e. a. Xuze, "A CNN Vehicle Recognition Algorithm based on Reinforcement Learning Error and Error-prone Samples," i *IOP Conference Series: Earth and Environmental Science*, 2018.

- [9] Y. Tian, S. Jana, K. Pei och B. Ray, "DeepTest: Automated Testing of Deep-Neural-Network-driven Autonomous Cars," i *ACM/IEEE 40th International Conference on Software Engineering*, Göteborg, 2018.
- [10] J. Yongwhan, K. Seonwook och Y. Kyongsu, "Surround Vehicle Motion Prediction Using LSTM-RNN for Motion Planning of Autonomous Vehicles at Multi-Lane Turn Intersections," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 1, s. 2-14, 2020.
- [11] F.-L. Vincent, P. Henderson, R. Islam, M. Bellemare och J. Pineau, *An Introduction to Deep Reinforcement Learning*, now publishers, 2018.
- [12] T. Okuyama, T. Gonsalves och J. Upadhay, "Autonomous Driving system based on Deep Q-learning," i *2018 International Conference on Intelligent Autonomous Systems*, 2018.
- [13] T. Tram, A. Jansson, R. Grönberg, A. Mohammed och J. Sjöberg, "Learning Negotiating Behavior Between Cars in Intersections using Deep Q-learning," i *21st International Conference on Intelligent Transportation Systems (ITSC)*, Maui, Hawaii, USA, 2018.
- [14] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao och K. Zieba, "End to End Learning for Self-Driving Cars," Holmdel, New Jersey, 2016.
- [15] P. Lin, "Why Ethics Matters for Autonomous Cars," i *Autonomous Driving: Technical, Legal and Social Aspects*, Springer Open, 2016, s. 69-79.
- [16] J.-F. Bonnefon, A. Shariff och I. Rahwan, "The social dilemma of autonomous vehicles," *Science*, 2016.
- [17] A. Hars, "Top misconceptions of autonomous cars and self-driving vehicles," *Inventivo Innovation Briefs*, Nuernberg, 2016.
- [18] A. Islam och I. S. Rashid, "Algorithm for Ethical Decision Making at Times of Accidents for Autonomous Vehicles," i *4th International Conference on Electrical Engineering and Information & Communication Technology*, 2018.
- [19] C. J. Gerdes och S. M. Thornton, "Implementable Ethics for Autonomous Vehicles," i *Autonomous Driving: Technical, Legal and Social Aspects*, Springer Open, 2016, s. 87-88.
- [20] S. Nyholm och J. Smids, "The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?," *Ethical Theory and Moral Practice*, nr 19, s. 1275-1289, 2016.
- [21] N. J. Goodall, "Ethical Decision Making During Automated Vehicle Crashes," *Transportation Research Record Journal of the Transportation Research Board*, vol. 2424, s. 58-65, 2014.

