Norsk filosofisk tidsskrift

Universitetsforlaget

VITENSKAPELIG PUBLIKASJON

Årgang 58, nr. 2–3–2023, s. 118–128
ISSN online: 1504-2901
DOI: https://doi.org/10.18261/nft.58.2-3.5

Moralsk ansvar for handlinger til autonome våpensystemer

Moral responsibility for the actions of autonomous weapon systems

Kjetil Holtmon Akø
Doktorgradsstipendiat, Institutt for religion, filosofi og historie, Universitetet i Agder
Kjetil Holtmon Akø forsker i sin doktorgrad i filosofi på kunstig intelligens og etikk med fokus rettet mot spørsmål om ansvar samt teknologibruk og -utvikling.
Kjetil.h.ako@uia.no

Sammendrag
Spørsmålet om hvem som bør holdes ansvarlig for handlingene til autonom teknologi er blitt aktualisert de siste tiårene. Utvikling innenfor kunstig intelligens har ført til mer kapable roboter som kan utføre kompliserte oppgaver på egen hånd. Dette har reist spørsmål om hvem som er ansvarlig for handlingene begått av slik teknologi da den opererer på en måte hvor det kan være vanskelig eller umulig å gripe inn og hindre uønskede handlinger. I denne artikkelen kritiserer jeg et argument fremmet av Lode Lauwaert som forsøker å plassere ansvaret for en krigsforbrytelse begått av et autonomt våpensystem hos offiseren.

Nøkkelord
Kunstig intelligens, KI etikk, autonom våpenteknologi, moralsk ansvar, kontroll

Abstract
In recent decades, the issue of responsibility for the actions of autonomous technology has become increasingly relevant. Advancements in artificial intelligence have resulted in highly capable robots capable of performing complex tasks independently. Consequently, questions have arisen regarding who should be held responsible for the actions carried out by such technology, especially when it operates in ways that make it challenging or even impossible to intervene and prevent undesirable actions. In this article, I will critique an argument put forth by Lode Lauwaert, which proposes placing responsibility for a war crime committed by an autonomous weapons system on the commanding officer.

Keywords
Artificial intelligence, AI ethics, lethal autonomous weapons system, LAWS, moral responsibility, control

1. Introduksjon
I mars 2020 ble stridende styrker tilknyttet Khalifa Haftar angrepet av autonome våpensystemer under en tilbaketrekning i Libya (FN, 2021). Ifølge rapporten fra FN er dette det første bekreftede tilfellet av et våpenangrep der beslutningen om å angripe spesifikke mål ikke ble kontrollert av en menneskelig operatør, noe som kan omtales som det første autonome våpendrapet. Det autonome våpensystemet som utførte angrepet var det tyrkisk-

produserte STM Kargu-2, som fungerer ved selvstendig å søke etter mål, deretter fly nært inntil målet, eksploderer og slynger prosjektiler rundt seg, uten å referere beslutningen til en operatør og be om godkjenning.

Gjør vi noen endringer i dette handlingsforløpet og forestiller oss at dette autonome våpensystemet angrep en ikke-stridende, så står vi ovenfor en mulig krigsforbrytelse. Dette reiser viktige spørsmål om hvem, om noen, som kan og bør holdes moralsk ansvarlig for en slik handling. Jeg er her interessert i spørsmålet om offiseren fortjener klander for handlingen til et autonomt våpensystem. Det vil si, jeg er opptatt av rettferdigheten av å klandre offiseren for en krigsforbrytelse begått av et autonomt våpensystem hvor offiseren ikke hadde mulighet til å kontrollere handlingsforløpet i sanntid.[1] Helt spesifikt vil jeg kritisere et argument som konkluderer at offiseren er moralsk ansvarlig og fortjener moralsk klander.[2]

Et argument som søker å opprette denne konklusjonen er blitt fremmet av Lode Lauwaert (2021). Hans argument er igjen et forsøk på å tilbakevise Robert Sparrows argument, som forsøker å vise at i) ingen kan holdes ansvarlig for krigsforbrytelser begått av autonome våpensystemer og ii) autonome våpensystemer er i strid med krigens folkerett (Sparrow, 2007). Ettersom mitt anliggende er spørsmålet om hvorvidt noen er moralsk ansvarlige for slike handlinger, ser jeg bort fra Sparrows andre argument.

Formålet her er ikke å forsvare Sparrows konklusjon, men å vise at motargumentasjonen til Lauwaert har svakheter og dermed ikke er et sterkt angrep på Sparrows posisjon dersom sistnevntes posisjon leses som et forsvar for urettmessigheten i å moralsk klandre offiseren. Videre, denne artikkelen begrenser seg til å omhandle moralteoretiske avveininger og argumenter om den moralske ansvarligheten til et bestemt individ blant flere og er dermed ikke ment som et innlegg henimot et oppdatert regelverk for krigføring. Der Sparrow og Lauwaert behandler moralsk og juridisk ansvar sammen, velger jeg å adskille dem og retter søkelyset på førstnevnte, da på en bestemt type klander. Dette innebærer ikke at jeg overser krigens regler som relevant ettersom disse utgjør regler og normer for stridende som er relevante for spørsmål om ansvar.

Noen presiseringer er på sin plass. Sparrows argument er nokså sterkt i den forstand at han konkluderer at ingen kan og bør holdes juridisk og moralsk ansvarlig for disse handlingene. Ifølge han fortjener ingen klander og må derfor ikke holdes ansvarlig. Slik jeg tolker Sparrow, er han opptatt av den type klander som er av en sterkere art (Sparrow, 2007, fn. 48). Argumenter som forsøker å opprette denne og lignende konklusjoner omtales på engelsk som *responsibility gaps*, på norsk *ansvarshull*, og knytter seg både til våpenteknologi og annen autonom teknologi.[3] Grovt sett argumenterer forkjemperne for at våre moralske ansvarspraksiser ikke klarer å ta høyde for denne typen teknologi, og det oppstår en uoverkommelig distanse mellom ønsket om å holde personer ansvarlig og rettferdigheten i å gjøre dette. Hos Sparrow oppstår denne distansen fordi ingen har tilstrekkelig kontroll over og kunnskap om enkelthandlingene til autonome våpensystemer (Sparrow, 2007, s. 69–71). Kontroll og kunnskap er stipulert som individuelt nødvendige

1. Dermed er jeg her ikke opptatt av spørsmålet om relevante individer tilfredsstiller betingelser for juridisk ansvar. Videre er oppmerksomheten rettet mot samtidig våpenteknologi da jeg nødig vil forskuttere hvilke teknologiske egenskaper som kan oppstå som følge av utvikling innenfor teknologifeltet.
2. I likhet med Sparrow og Lauwaert har jeg valgt å ikke skille mellom ulike offiserer i kommandokjeden. Videre arbeid bør skille mellom militært personell, som på ulikt vis bidrar til godkjenning og bruk av autonome våpensystemer. Andre aktører som utviklere, produsenter og politiske beslutningstakere vil være relevant å trekke inn for å avgjøre ansvars- og skyldspørsmålet.
3. Se Danaher (2016); de Jong (2020); Matthias (2004) for ulike varianter av disse. Matthias sin artikkel regnes for å introdusere ansvarsproblematikken i konteksten av autonome teknologier.

og i fellesskap tilstrekkelige betingelser for å rettmessig kunne holdes moralsk ansvarlig for en handling.

Videre forstår Sparrow autonomien til disse våpensystemene i en sterk forstand: de kan fatte beslutninger på selvstendig vis, de er intelligente i måten de tilnærmer seg sine mål og oppgaver, og de kan lære fra sine omgivelser og med dette gjøre endringer i sine interne sammensetninger (Sparrow, 2007, s. 65). Dette gjør dem åpenbart uforutsigbare og vanskelige å kontrollere ettersom de kan forandre reglene for sin egen operasjon. I tillegg er de tiltenkt å operere fullstendig uavhengig av operatøren. Dermed kan for eksempel ikke operatøren være sikker på om våpensystemet angriper det intenderte målet og fortjener med det ikke klander når det begår en krigsforbrytelse (Sparrow, 2007, s. 70).

I likhet med Lauwaert tror jeg imidlertid vi bør være forsiktige med å anta at autonomien til slike maskiner hviler på en evne til å lære samtidig som de er i operasjon. En slik evne vil lede til stor usikkerhet om hva de kan gjøre, men det synes ikke å være nødvendig for usikkerhet som sådan. Jeg kommer tilbake til dette senere i teksten, men før den tid skal vi se på Lauwaerts påstand. Felles for Sparrow, Lauwaert og meg er at situasjonen som opptar oss, er den hvor autonome våpensystemers beslutninger og handlinger ikke kan kontrolleres og stoppes i sanntid. Operatøren er med andre ord «off-the-loop» og kan ikke gripe inn.

2. Lauwaert og ansvar for autonome våpensystemer

Lauwaert ber oss vurdere det følgende scenarioet: «Suppose that the army decide to utilize LAWS [lethal autonomous weapons systems]. The robot decides on its own to fly out, chooses the target and finally kills a civilian of the hostile camp.» (Lauwaert, 2021, s. 1006). Dette er foranlediget av en ordre fra en offiser, men selve sekvensen beskrevet er utført utelukkende av det autonome våpensystemet. Lauwaert følger dermed Sparrow et stykke på vei når det gjelder autonomien til roboten, men avstår fra å inkludere evnen til å lære. Det vesentlige er denne teknologiens evne til å operere uavhengig av en operatør og at denne evnen er delvis bestående av moderne kunstig intelligente løsninger (Lauwaert, 2021, s. 1002–1003). Han skriver at denne formen for kunstig intelligens ikke er *good-old-fashioned AI*.[4] Dette siste skal tolker jeg dithen at slike roboter består av maskinlæringsalgoritmer av en spesifikk type, uten at evnen til å lære fastholdes når de er i bruk. Det vil med andre ord si at de består (delvis) av algoritmer som er blitt trent opp, men hvor treningen ikke fortsetter når roboten er i bruk. Ettersom Lauwaert karakteriserer disse maskinlæringsalgoritmene som ugjennomtrengelige for vår forståelse (vi vet ikke hvordan de prosesserer informasjon og kan ikke med sikkerhet vite det), er det nærliggende å tenke at han har dype nevrale nettverk i tankene.

Ettersom dette er en mulig krigsforbrytelse, er det ønskelig å avklare hvem som er moralsk ansvarlig for handlingen slik at de rette kan holdes ansvarlig. Våpensystemet selv er ifølge Lauwaert ikke en mulig kandidat, da det mangler evnen til å oppleve klander som noe smertefullt. Klander er ifølge ham nødvendigvis knyttet til smerte, og motsatt er prisverdighet for gode handlinger nødvendigvis opplevd som godt (Lauwaert, 2021, s. 1007). Muligheten for å oppleve klander og ros anses med andre ord som en forutsetning for å kunne bli holdt moralsk ansvarlig for ens handlinger, konsekvenser og utelatelser.

Han stipulerer videre tre andre nødvendige betingelser. En aktør S er ansvarlig for x (der x kan være en handling, konsekvens eller utelatelse) dersom:

4. Good-old-fashioned AI (GOFAI) brukes ofte som et samlebegrep for kunstig intelligens utviklet gjennom symbolsk programmering og eksplisitte regler og brukes som en kontrast til moderne maskinlæringsalgoritmer der algoritmen selv lager representasjoner og finner sammenhenger ved å lære fra datasett.

i. S er kausalt tilknyttet x på en slik måte at S sitt bidrag forklarer hvorfor og hvordan x skjedde (den kausale betingelsen),
ii. S besitter intellektuell autonomi, det vil si, S kan veie grunner for og imot en handling og fatte en beslutning på bakgrunn av dette (betingelsen om intellektuell autonomi) og
iii. S besitter moralske og faktuelle oppfatninger om x, det vil si, S vet om x er tillatt, forbudt eller påbudt og vet om x kan følge kausalt av hens handling (den epistemiske betingelsen) (Lauwaert, 2021, s. 1007).

Det er verdt å merke seg her at Lauwaert ikke anser kontroll som en nødvendig betingelse for å være moralsk ansvarlig for x. Til støtte for påstanden om ikke-nødvendigheten av kontroll ber han oss vurdere to situasjoner med epileptiske anfall. I det første tilfellet har person A ingen sykdomshistorikk med epilepsi og får et anfall samtidig som hen er ute og kjører en bil. I det andre tilfellet har person B en slik historikk og vet om risikoen for å få et anfall samtidig som hen kjører. I begge tilfeller er valget begrunnet av et ønske om å unngå en regnskur. I begge tilfeller blir en syklist påkjørt som følge av et epileptisk anfall. Lauwaerts klare intuisjon er at kun sistnevnte individ vil vi holde ansvarlig for ulykken (Lauwaert, 2021, s. 1008). Samtidig, i begge tilfeller er det et fravær av kontroll over bilen som fører til ulykken. Lauwaert konkluderer fra dette at kontroll ikke er nødvendig for å holde personer moralsk ansvarlige dersom de tre andre betingelsen er tilfredsstilt.

Videre avgrenser Lauwaert sin søken etter en ansvarlig part til de direkte involvert i bruken av teknologien. Det er uklart akkurat hvorfor Lauwaert oppretter denne begrensningen, men det virker å forklares ved en henvisning til hvilken aktør som sist bidrar til ugjerningen, altså den siste utløsende faktoren. Dette hviler nok på tanken om at brukere av teknologien tar over ansvaret for den spesifikke bruken av teknologien, og at utviklere og produsenter kun er ansvarlig dersom teknologien ikke fungerer som informert. Man kan forestille seg, slik Lauwaert antar, at risikoen for slike ugjerninger er informert om på forhånd (Lauwaert, 2021, s. 1008). Dermed aksepterer brukeren risikoen dersom hen velger å benytte seg av teknologien.[5]

På basis av disse betingelsene og situasjonsbeskrivelsen, argumenterer Lauwaert at offiseren er moralsk ansvarlig for handlingen til maskinen. Han oppsummerer:

> Let us assume that the commander knows the principles of just war: she or he knows that it is forbidden to intentionally kill civilians and to use excessive force. And let us also assume that the war crime is not an unforeseen consequence: the commander knows that there is a risk that the robot's algorithm will result in the command to fire at civilians or to use disproportionate force. Well, when this actually happens, you can hold the commander responsible and call her or him to account […] The reason is that the commander (fully aware of the risk of deaths) decides to use the LAWS, and that the decision is causally linked to the deaths of the civilians. (Lauwaert, 2021, s. 1008)

I lys av betingelsene ovenfor får vi:

Offiseren er ansvarlig for det autonome våpendrapet av en sivil (krigsforbrytelsen) dersom:

i. Offiseren er kausalt tilknyttet krigsforbrytelsen som en utløsende faktor,
ii. offiseren veide grunner for og imot å ta i bruk det autonome våpensystemet, og
iii. offiseren vet dette er en krigsforbrytelse og moralsk galt, samt at denne handlingen er et sannsynlig utfall av bruken.

5. Flere forfattere innenfor dette temaet fordrer samme tanke, deriblant Danaher (2016); Matthias (2004); Oimann (2023).

Dermed er spørsmålet om hvem som fortjener klander og muligens straff avklart i konteksten av krigsforbrytelser begått av autonome våpensystemer. Lauwaert skriver «At best, the thought experiment [den epileptiske sjåføren] shows that Sparrow's second premise is not supported by the argument that control is necessary for responsibility.» (Lauwaert, 2021, s. 1008). Ettersom kontroll ikke er en nødvendig betingelse, så spiller det ingen rolle om operatøren har mulighet til å gripe inn og avbryte handlingen. Det som er avgjørende for spørsmålet om moralsk ansvar er hvorvidt offiseren foretok et uforsvarlig valg som forårsaket hendelsen. Vedkommende som vel vitende om risikoen for et uønsket utfall handler uforsvarlig i lys av dette. Det er en overveid beslutning om å trosse risikoen for en mulig ulykke som konsekvens av hens handling som for Lauwaert innebærer at vedkommende er moralsk ansvarlig i denne situasjonen (Lauwaert, 2021, s. 1008).

Mitt anliggende i den påfølgende delen er å problematisere to sider ved Lauwaerts posisjon. Først, er dette et klart tilfelle av en uforsvarlig handling fra befalet sitt side? Altså, skal vi godta antagelsen at krigsforbrytelsen er en forventet konsekvens? Mye av styrken i argumentet til Lauwaert hviler på nettopp denne antakelsen.

Deretter, Lauwaert ser ut til å innrømme kontroll som vesentlig for spørsmål om ansvarlighet. Selv om han ikke bruker begrepet om kontroll og benekter nødvendigheten av dette, så er det elementer i hans argument og eksempler som avslører kontroll. Denne formen for kontroll er relevant for å besvare spørsmålet om offiseren fortjener en form for klander for ugjerningen, men ikke slik Lauwaert ser for seg. Den relevante formen for kontroll er den som legitimerer sterkere former for klander, og denne er ikke til stede hos Lauwaert. Slik sett treffer ikke Lauwaerts kritikk av Sparrow særlig godt ettersom sistnevnte er interessert i spørsmålet om offiseren fortjener en sterkere form for klander.

3. Akseptabel og uakseptabel risiko

Det er liten tvil at risikovillig oppførsel til tider kan kvalifisere som oppførsel for hvis en person fortjener klander, men spørsmålet er hvorvidt krigsforbrytelsen i eksempelet ovenfor tilsvarer en urimelig høy risiko og dermed er klanderverdig oppførsel. Er enhver kunnskap om risiko for uønsket konsekvens nok til å anspore klanderverdighet? Hensynsløshet betyr i denne sammenheng å ta en uforsvarlig risiko. For å vurdere styrken i Lauwaerts påstand, må vi i så henseende forstå alvorligheten i risikoen og dens sannsynlighet.

Som sagt, Lauwaert oppfatter det som åpenbart at den epileptiske sjåføren er hensynsløs; hen godtar en uberettiget risiko ved å fatte beslutningen om å sette seg inn i bilen og kjøre av gårde vel vitende om at hen når som helst kan få et epileptisk anfall. Dermed kan vi tenke oss at hen bryter en forpliktelse, muligens å ikke utsette andre for unødvendig skade, fordi grunnen for å handle slik ikke er god. Med andre ord, vedkommende feiler i å forstå at de tungtveiende grunnene for at hen ikke skal sette seg bak rattet og kjøre.

Man kan òg konstruere et eksempel som ligner, men hvor konklusjonen om klanderverdighet synes å peke i en annen retning. Forestill deg så at du er ute og kjører en solskinnsdag. Som kjent kan sterkt solskinn til tider få en til å nyse. Du vet like godt at du ikke nyser hver gang det er sterkt solskinn, så tanken slår deg enkelt og greit ikke. Likevel, som en rasjonell overveier av informasjon er det uproblematisk å komme til den slutning at dersom du nyser mens du kjører, så kan det gå riktig galt, fort. Slik går det. Du nyser, mister midlertidig kontrollen over bilen og treffer en syklist. Er vi i dette tilfellet like tilbøyelige til å klandre sjåføren for det som skjedde, slik Lauwaert hevder? I både dette eksempelet og med den epileptiske sjåføren lider personene av en umiddelbar og ufrivillig kroppslig reaksjon som resulterer i manglende kontroll over bilen.

Dette kan problematiseres videre i lys av andre medisinske tilstander. Gitt utbredelsen av akutte hjerteinfarkt, hjerneslag og andre sykdommer som er utbredt i befolkningen, og den påfølgende statistiske sannsynligheten for at et individ kan lide av plutselige symptomer, så vil det ifølge Lauwaerts betingelser åpnes for mer rettmessig klander. Spørsmålet det koker ned til, er hvor sterk sannsynligheten er for at noe kan gå galt, må overveies av et individ før det vil være rettferdig å utsette individet for smertefull klander.

Argumentet til Lauwaert hviler i stor grad på antakelsen om at utfallet er tilstrekkelig forventet til å regnes som uforsvarlig risikotagning. Uten denne antakelsen blir det vanskeligere å konkludere at oppførselen er klanderverdig ettersom kunnskapen om risiko er fraværende. Det er samtidig grunn til å stille spørsmål ved hvor stor risikoen for uønskede utfall må være for at oppførsel skal anses som uforsvarlig. Med andre ord, når bør man vite om risikoen for uønsket utfall er stor nok til å unnlate å handle? Skal vi da tro Lauwaert på at kunnskap om mulige uønskede konsekvenser av ens handlinger er tilstrekkelig til å handle uforsvarlig, uavhengig av hvor sannsynlig det er at de inntreffer? Som jeg antydet ovenfor, ser en slik tilnærming ut til å gi oss dommer som fremstår som problematiske i lys av omfanget av berettiget klander den tillater. Ved å akseptere premissene til Lauwaert åpnes det for fortjent klander i en vid rekke tilfeller der personer ufrivillig mister kontroll.

Grovt sett er klander berettiget dersom vi har å gjøre med en moralsk aktør (et passende objekt for våre reaksjoner), en ugjerning, og at den moralske aktøren er knyttet til ugjerningen på en passende måte. Jamfør betingelsene til Lauwaert må «passende måte» forstås som aktøren er en utløsende årsak hvor handlingen eller effekten er kjent for aktøren som moralsk uønsket.

Enkelte konsekvenser av ens handlinger vil imidlertid være så uventede at å holde individer ansvarlig for slike konsekvenser innebærer å stille et meget strengt krav til dem. Strengt fordi rettmessig klander avhenger av klanderverdighet, og dette styres av rettferdighetsbetraktninger om individets mulighet til å unngå den smerten som kan følge med det å bli klandret. Lauwaert synes å godta denne begrensningen når det er rettmessig å klandre. Han anerkjenner at enkelte faktorer spiller inn på spørsmålet om klanderverdighet. Han skriver «Those who do not know that an action is wrong cannot be held responsible for it (unless you are guilty of ignorance).» (Lauwaert, 2021, s. 1007) og det samme for konsekvenser av ens handlinger. Den underliggende tanken er: ettersom klander er smertefullt, så må klanderen fortjenes for at den skal være rettferdig. Er man uvitende om noe er moralsk tillatt eller forbudt eller om konsekvensene som følger, så vil det være vanskelig å følge de moralske normene på bakgrunn av de rette grunnene. Fra perspektivet til det handlende individet så handler man i god tro om at man ikke bryter en moralsk norm, og det blir vanskelig å handle annerledes.

En slik rettferdighetsbetraktning forklarer også noe av det historiske fokuset på kontroll i ansvarsspørsmål. Et individ som frivillig og vitende begår en handling og vet hvilke konsekvenser som sannsynligvis følger, har kontroll over det hen gjør og det som kommer som en effekt av hens handlinger. Dermed kunne individet òg valgt å ikke handle slik. Siden det da var opp til individet hva hen gjorde, så er klanderen fortjent. Det er ikke urettferdig å bli klandret for ugjerninger man med viten og vilje gjør, men det er en urett å klandres for handlinger og konsekvenser som er uunngåelige og som man ikke kontrollerer dersom klanderen medfører smerte. Uvitenhet gjør det som sådan vanskelig å unngå å handle galt da man mangler relevant informasjon som ville informert ens beslutning. Slik sett kan rettferdighetsbetraktningen formuleres som en mulighet til å unngå smertefull klander. Sparrows bekymring knyttes til dette, om enn ikke skrevet ut i særlig grad (Sparrow, 2007, s. 69–71 og fn. 48).

Men ifølge Lauwaert er slik kontroll ikke nødvendig. I stedet er klander fortjent dersom individet vet om tillateligheten av handlingen og dens konsekvenser, og at denne vurderingen og handlingsforløpet på en passende måte springer fra individet. At dette ser ut til å utgjøre en form for kontroll kommer jeg tilbake til.

Jeg er ikke helt overbevist om at kunnskap om enhver risiko gjør det rettferdig å holde offiseren ansvarlig. Først og fremst er det ikke overbevisende at kunnskap om en hvilken som helst grad av risiko for feilhandlinger tilsier klanderverdighet i alle tilfeller. Hvis alt går bra i 99 prosent av tilfellene, er vi enige i at den ene prosenten av tilfellene der noe går galt også er klanderverdige i kraft av vissheten om at dette kan bli tilfelle på et tidspunkt? Vi vil trenge et avgrensningsprinsipp som tydeliggjør hvilket risikonivå som er uakseptabelt i tilfeller hvor muligheten til å unngå ugjerninger og uønskede konsekvenser er fraværende, som i tilfeller der befalet til et autonomt våpensystem ikke kan gripe inn og hindre visse handlinger.

En mulig måte å skille mellom de konsekvensene som faller utenfor det vi anser som rimelig å vurdere, er å se om agenten mislyktes i å oppfylle sine epistemiske forpliktelser i en bestemt situasjon. Vi har noen forventninger til hva agenter bør vurdere før de handler, hvilken informasjon de bør søke å oppnå og hvor mye innsats som bør brukes på å prøve å finne riktig informasjon, slik at de kan få kunnskap om skaden som kan følge av deres handlinger eller på annet vis bli oppmerksom på tillateligheten av deres handlingers konsekvenser. Det epistemiske kravet kan da tas for å si at det bare er rimelig å forvente at agenter forutser de konsekvensene hen kunne ha oppdaget som sannsynlige hendelser dersom agenten hadde oppfylt de epistemiske forpliktelsene. Og med tanke på at noen konsekvenser, selv om de kunne bli oppdaget, ville være usannsynlige hendelser hvis de skulle inntreffe, kan det her være rom for å benekte Lauwaerts argument.

Faktisk kan vi tenke oss at epistemiske forpliktelser vil begrense individets tiltro til sannsynligheten for enkelte konsekvenser. Disse konsekvensene, selv om mulige effekter av ens handling, vil være usannsynlige konsekvenser. Dermed bør vi ikke forvente at aktører forut for enhver handling hvor det foreligger en usannsynlig mulighet for uønskede konsekvenser skal bli klar over disse konsekvensene og deretter avstå fra å handle på basis av denne tanken.

Troen på at et autonomt våpensystem vil begå en krigsforbrytelse er et utilstrekkelig grunnlag for offiserens klanderverdighet, med mindre risikoen for at det vil skje er tilstrekkelig høy. Dersom oddsen ikke er sterkt i favør for at en krigsforbrytelse kan oppstå, så bør vi ikke anse dette som et tilfelle av klanderverdig oppførsel. Og en slik odds virker ikke å være særlig høy.[6]

Utenfor lovens rammeverk virker vi ikke tilbøyelige til å holde individer ansvarlig for alle uhell tilknyttet bruk av teknologi dersom disse ikke er forårsaket av uaktsomhet, hensynsløshet eller andre klanderverdige brudd på forpliktelser. Uventede og utilsiktede effekter ser ut til å være karakteristiske for større komplekse systemer med mange samvirkende

6. Vi kan forestille oss at vedkommende har brukt våpensystemet over lengre tid uten å oppleve noen problemer, og som en følge av dette underholder muligheten i enda mindre grad enn det feilprosenten tilsier, uten at dette er noe hen er klar over. Uten en oppmerksomhet på denne feilvurderingen kan ikke vedkommende starte gjenopprettingen av den «riktige» innstillingen ovenfor teknologibruken. Dette innebærer derimot ikke at andre typer ansvar, som juridisk eller rolleansvar, ikke kan trekkes inn, men det spiller en vesentlig rolle for spørsmålet om moralsk ansvar og klanderverdigheten – særlig klander som rettferdiggjør sterkere sanksjoner og reaksjoner. Videre, denne antakelsen hviler på prosessene for verifisering og validering av teknologien som den går gjennom før den godkjennes i operative tjenester, samt at feilprosenten er lav på bakgrunn av disse prosessene. Jeg antar videre at feilprosenten i forbindelse med autonome våpensystemer vil være tilsvarende lik som for andre våpensystemer og missiler.

deler, og det kan være upraktisk og uoppnåelig å forvente at slike konsekvenser ikke av og til vil inntreffe med autonom teknologi. Men det er uklart hvordan det er rettferdig å tilskrive brukere av teknologien ansvaret for teknologisk ugjerning når denne feilhandlingen ikke fullt ut kan forklares av brukernes oppførsel, og at det heller ikke er en forseelse som vi med rimelighet kan forvente disse aktørene å forutse.

Det problematiske er antakelsen at en slik krigsforbrytelse alltid må anses som en forventet konsekvens ved bruken av slike våpen. Særlig problematisk virker antakelsen om forventet konsekvens å være dersom det autonome våpensystemets evne til å kategorisere objekter i miljøet består av dype nevrale nettverk, som er maskinlæringsteknikken Lauwaert virker å ha i tankene. Denne maskinlæringsteknikken anses for å være ugjennomsiktig. De matematiske kalkulasjonene som foregår inni dem, er foreløpig for kompliserte til at vi har en god forståelse av det som foregår (Barredo Arrieta et al., 2020, s. 83; Cappelen & Dever, 2021, s. 10; Rudin, 2019, s. 207). Dette gjør det vanskelig å fullt ut verifisere hvordan input leder til output, noe som skaper et usikkerhetsmoment og øker sjansen for uintenderte og uforventede konsekvenser.[7]

Når et autonomt våpensystem bestående av slik teknologi benytter denne teknikken for å kategorisere omgivelsene sine, kan det lede til tilfeller der systemet feilaktig tildeler en kategori til et objekt. Dersom ingen har mulighet til å overvåke og gripe inn, kan en slik feilkategorisering bli bearbeidet innad i systemet og lede til en handling som ikke er ønsket. Våpensystemet vil simpelthen handle på feil grunnlag og velger en handling som ikke stemmer overens med det omgivelsene og de intenderte reglene tilsier er den korrekte. Slik feilkategorisering er allerede rapportert å ha funnet sted i enkelte biler som benytter denne teknologien.[8] Det er viktig å merke seg at teknologien her ikke skaper nye regler og handlinger på basis av informasjon i omgivelsene, men at den feiltolker omgivelsene og sammen med de programmerte reglene, velger feil handling på feil sted.

En slik antakelse virker nå å være problematisk i tilfeller ved bruk av autonome våpensystemer. I disse tilfellene er det ikke en klar hensynsløs og uforsvarlig oppførsel fra offiseren ettersom de ikke bryter forpliktelser til å holde seg informert da krigsforbrytelsen som konsekvens er lite sannsynlig og vanskelig å forutse.

Dette blir særlig påfallende dersom vi beholder Lauwaerts premiss at slik teknologi fungerer som informert, altså at brukeren er informert om at teknologien har en iboende risiko for å begå fatale feil, og at de som utvikler og produserer ikke er klanderverdige for teknologien de tilbyr. En slik tankegang ser ut til å bero på en instrumentell forståelse av teknologi, der teknologi oppfattes som et middel til et mål. Utvikleren og produsenten er kun ansvarlig dersom teknologien fungerer som et middel til et mål slik den var ment å fungere (og hvis vi antar at disse ikke er forbudte). Fungerer teknologien som den er informert om, så faller alt ansvar for eventuelle uhell på brukeren. En slik tankegang overser at dårlige teknologiske løsninger kan spores til de som utvikler og produserer teknologien.

I det foregående har jeg forsøkt å vise at mengden rettmessig klander ifølge Lauwaerts stipulerte betingelser leder til mer klander enn det som virker rettmessig, samt at den teknologien Lauwaert har i tankene undergraver tanken om krigsforbrytelser som forventede konsekvenser ved bruk. Derfor vil det være urettferdig ovenfor offiseren å utsette hen for den smerte som kan komme i form av klander, særlig den videre straffen som kan utledes av å bli holdt ansvarlig.

7. Se Carabantes (2020) for en kritisk gjennomgang av forsøk på å forstå dype nevrale nettverk.
8. https://www.businessinsider.com/video-tesla-autopilot-appears-to-confuse-horse-drawn-carriage-truck-2022-8?r=US&IR=T

4. Lauwaert og kontroll

Lauwaert påstår at kontroll ikke er nødvendig for ansvar, men dette er feil. I det ovennevnte sitatet hentet fra artikkelen til Lauwaert benekter han eksplisitt at kontroll er en nødvendig betingelse for å bli hold ansvarlig. Dersom kontroll ikke er nødvendig slik Lauwaert hevder, er det vanskelig å forstå det han skriver at en nødvendig del av å kunne bli holdt moralsk ansvarlig for en konsekvens innebærer at aktøren «spiller en kausal rolle» i hendelsene som fører til konsekvensen. Det er nærliggende å tolke dette som at aktøren handler slik at konsekvensen forklares avgjørende av handlingen. Men hvis dette er tilfellet, så må Lauwaert innrømme tilstedeværelsen av kontroll som nødvendig for moralsk ansvar. Kontroll i form av at handlingen springer ut av aktørens egen evne til å vurdere grunner for og imot en handling, kunnskap om hvilke konsekvenser som er sannsynlig å være en effekt av handlingen, samt den kausale sammenhengen mellom handling og konsekvens.

Kontroll er nødvendig for ansvar dersom vi har å gjøre med reaksjoner av en sterkere art på grunn av uretten ved å straffe noen som mangler kontroll, og som samtidig ikke klanderverdig ga fra seg kontroll ved et tidligere tidspunkt. Lauwaert burde derfor ha skrevet at direkte kontroll ikke alltid er nødvendig for å kunne holdes ansvarlig. Gitt dette kan påstanden hans omformuleres: offiseren utøver tilstrekkelig indirekte kontroll over krigsforbrytelsen til å rettferdiggjøre sterkere typer klander som kan innebære at den klandrede part opplever smerte. Det er den den frivillige og veloverveide beslutningen om å bruke våpensystemet og den påfølgende effekten av tap av direkte kontroll som er startpunktet for å holde offiseren ansvarlig.

Spørsmålet om kontroll og hvor godt Lauwaerts innvending mot Sparrow treffer, henger sammen med hvorvidt vi har med en uforsvarlig risikotakning å gjøre. Sparrow og Lauwaert skiller lag i hvordan de tenker om kontroll. Sparrow forstår kontroll, i forbindelse med ansvarsspørsmålet, som evnen og muligheten til å påvirke et utfall i sanntid. Lauwaert derimot, har en bredere forståelse ved at å ha kontroll over et utfall, så er det tidvis tilstrekkelig med en visshet om hva som følger av ens handlinger, uavhengig av om man senere kan endre disse følgene. Dette innebærer at kontroll over et utfall kan spores bakover i tid, selv om individet i øyeblikket hvor situasjonen oppstår ikke kan foreta seg noe for å endre utfallet. Derfor er sannsynligheten for følgene av ens handlinger vesentlig for å avgjøre hvorvidt man har klanderverdig kontroll over det som skjer, og derfor er kunnskapsbetingelsen sentral for spørsmålet om klander. Slik sett tillater Lauwaerts begrep om ansvar at man kan være indirekte ansvarlig ved å spore klanderverdigheten til et tidligere tidspunkt der man hadde kontroll. På bakgrunn av dette er det høyst relevant å få klarhet i hvorvidt risikoen som tas av offiseren er moralsk forsvarlig.

Relevant fordi klanderverdig indirekte kontroll hviler på en hensynsløs eller uaktsom tilnærming til følgende av ens handlinger. Dette innebærer ikke at sikker kunnskap må være til stede for at et individ skal rettmessig moralsk klandres, men konsekvensene må rimelig kunne forutses.

Det virker rimelig å anta at uaktsomhet ikke er relevant i denne sammenheng gitt opplæringen militært personell mottar i forbindelse med våpentrening. Spørsmålet er dermed om offiseren opptrådde hensynsløst da hen ga fra seg muligheten til å intervenere i våpensystemets målutvelgelse, og svaret på dette spørsmålet avhenger av i) sannsynligheten for feil målutvelgelse og ii) relevante deler av krigens regler som proporsjonalitetsprinsippet og doktrinen om dobbel effekt. I den foregående delen ga jeg noen grunner til å tro at offiseren ikke har handlet uforsiktig.

Videre, det er verdt å merke seg at Lauwaerts benektelse av kontroll og det han skriver om attributivt ansvar og dets sammenheng med evaluerende dommer (Lauwaert, 2021, s.

1003) gir grunn til å tolke han som å fremme en forståelse av ansvar som «attributability».[9] I litteraturen om moralsk ansvar er det vanlig å skille mellom former for klander og evalueringer og grunnlaget som gjør disse passende. Det er utbredt å forstå moralsk ansvar som bestående av to typer reaksjoner: «attributability» og moralsk ansvar som «accountability». For de som fordrer en todelt forståelse av moralsk ansvar, blir førstnevnte vanligvis forbeholdt karaktervurderinger, mens sistnevnte er forbeholdt reaktive holdninger og rettferdiggjør potensielt skadelige opplevelser på vegne av den skyldige parten (f.eks. å få sine interesser diskreditert eller andre moralske sanksjoner). Sistnevnte, gitt muligheten for skadelige reaksjoner, er vanligvis ansett å kreve en sterkere kontrollbetingelse enn førstnevnte (Watson, 1996). Siden skade er involvert, blir hensyn til rettferdighet sentralt grunnet urettferdigheten ved å skade noen som ikke kunne unngå å handle på en bestemt måte eller som ikke var uforsvarlig eller uaktsomt informert om konsekvensene av sin atferd. Den typen ansvar som opptar Sparrow er nettopp den som rettferdiggjør slike sterke reaksjoner, og den vil derfor kreve en strengere kontrollbetingelse som berettiger denne type reaksjoner ovenfor individet som klandres.

Dersom det er riktig å tolke Lauwaert som å foreslå at situasjonen med offiseren er attributivt ansvarlig for krigsforbrytelsen, så tilkjennegir selv dette en form for kontroll, selv om Lauwaert ikke tenker på det slik. Men selv om det finnes en form for underliggende kontroll hos Lauwaert (det riktige forholdet mellom handling og intensjon, grovt sett), så er ikke dette den kontrollen som er relevant for å besvare spørsmålet om sterkere former for reaksjoner, deriblant straff som går utover fordømmelse av en persons karakter, reist av Sparrow. Den kontrollen som er relevant for dette, er den hvor individet har muligheten til å unngå handlinger som kan utgjøre en basis for disse sterke reaksjonene. Og der offiseren har evnen til dette i den forstand hen kan velge å ikke benytte seg av dronen, så mangler offiseren noe sentralt som spiller inn på hens evne til å benytte seg av denne muligheten: en rimelig oppfatning om at noe galt kan skje ved bruken av teknologien. Slik sett er det vanskeligere å se at offiseren har den kontrollen som er nødvendig.

5. Konklusjon

En klar grensesetting for risikotakning er utenfor omfanget til denne artikkelen, men jeg har forsøkt å problematisere det Lauwaert anfører ved å vise til at det kan gi bunn for et utvidet omfang av rettmessig klander. I forbindelse med den epileptiske sjåføren kan det stilles spørsmål ved hvorvidt det er sjåføren som godtar risikoen for et nytt anfall dersom vi antar at hens mulighet til å kjøre er godkjent av en lege og følger gjeldende lovgivning på området. På liknende vis kan vi anta at offiseren forholder seg til beslutninger fra øvre hold i hierarkiet samt andre offentlige instanser. Dersom vurderinger fra eksperter og andre beslutningstakere skal gi grunnlag for klanderverdig oppførsel, så kan det innebære at omfanget av rettmessig klander vil gå på bekostning av rettferdighetsbetraktninger ovenfor enkeltindivider. Det vil i praksis innebære at enkeltindivider ikke kan stole på vurderingene gjort av eksperter, men må ensidig akseptere risikoen på egen hånd og kontinuerlig foreta selvstendige vurderinger.

Det ser derfor ikke ut til at offiseren begår et klanderverdig normbrudd. Risikoen er antakeligvis lav, gitt godkjenningen av våpensystemet som foreligger. Det er heller ikke åpenbart at dette er et normbrudd gitt krigens regler, da særlig med tanke på doktrinen om dobbel effekt – muligens proporsjonalitetsprinsippet. Selv om offiseren har kontroll over

9. Se Talbert (2019) for en gjennomgang av standardposisjoner i feltet.

handlingen som igangsetter våpensystemet, så er det ikke klart at risikoen for feilutvelgelse av mål er tilstrekkelig stor til å utgjøre et uforsvarlig valg. En uforsvarlig beslutning innebærer at det er en rimelig forventning at utfallet av beslutningen kan bli annerledes enn intendert, og rimeligheten av dette vil avhenge av feilprosenten og muligens offiserens erfaring med bruk av det samme våpensystemet ved tidligere anledninger.

Dette innebærer dog ikke at andre aktører ikke kan være ansvarlige for det som skjedde, men jeg har valgt å ivareta Lauwaerts antakelse at offiseren er dit vi bør rette blikket. Politiske beslutningstakere, utviklere og produsenter er relevante aktører vi kan rette blikket mot i vår leten etter hvem som kan holdes ansvarlig.

Referanser

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., . . . Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Cappelen, H. & Dever, J. (2021). *Making AI Intelligible: Philosophical Foundations*. Oxford University Press.

Carabantes, M. (2020). Black-box artificial intelligence: an epistemological and critical analysis. *AI & SOCIETY, 35*(2), 309–317. https://doi.org/10.1007/s00146-019-00888-w

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology, 18*(4), 299–309. https://doi.org/10.1007/s10676-016-9403-3

de Jong, R. (2020). The Retribution-Gap and Responsibility-Loci Related to Robots and Automated Technologies: A Reply to Nyholm. *Science and Engineering Ethics, 26*(2), 727–735. https://doi.org/10.1007/s11948-019-00120-4

FN. (2021). *Final report of the Panel of Experts on Libya established pursuant to Security Council resolution 1973 (2011)* (S/2021/229). Retrieved from https://undocs.org/Home/Mobile?FinalSymbol=S%2F2021%2F229&Language=E&DeviceType=Desktop&LangRequested=False

Lauwaert, L. (2021). Artificial intelligence and responsibility. *AI & SOCIETY, 36*(3), 1001–1009. https://doi.org/10.1007/s00146-020-01119-3

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

Oimann, A.-K. (2023). The Responsibility Gap and LAWS: a Critical Mapping of the Debate. *Philosophy & Technology, 36*(1), 3. https://doi.org/10.1007/s13347-022-00602-7

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy, 24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Talbert, M. (2019). Moral Responsibility. I E.N. Zalta (Red.), *Stanford Encyclopedia of Philosophy (Winter 2019 edition)*.

Watson, G. (1996). Two Faces of Responsibility. *Philosophical Topics, 24*(2), 227–248. Retrieved from www.jstor.org/stable/43154245