Filosofiska aspekter på autonoma system

av Linda Johansson

Résumé

The use of autonomous military systems is increasing, and so is the development towards increasing autonomy in these systems. Before calling for prohibitions on even *developing* autonomous systems, which was suggested at the first multilateral talks on ”killer robots” held in Geneva 2014, it is important to clarify and look more closely at the philosophical aspects on these systems. This paper is an attempt to do so by looking at ethical issues on the use of autonomous systems, but also looking at philosophical issues such as responsibility, which involves philosophical areas such as free will and the philosophy of mind. The paper is a revised version of a talk held at the Royal Swedish Academy of War Sciences in January 2015.

ANVÄNDANDET AV AUTONOMA militära system ökar och utvecklingen går även mot ökad autonomi. Robotar som själva väljer och beskjuter mål kan vara verklighet om tjugo till trettio år. Detta hävdas av organisationen *Human Rights Watch* (HRW), med hänvisning till militära experter. Våren 2014 samlades 87 länder i Genève för att genomföra de första multilaterala samtalen om ”mördarrobotar”. Under samtalen framkom förslag på att faktiskt *förbjuda* inte bara tillverkandet och användandet av autonoma krigsrobotar, utan även *utvecklandet* av sådana. Frågan är emellertid om förbud är en förnuftig väg eller om man ska gå mer varsamt fram. Syftet med denna text är att genom en systematisering av de filosofiskt mest intressanta frågorna rörande militära robotar – för att både belysa och nyansera – utgöra ett litet bidrag till det sistnämnda synsättet. Texten är en omgjord version av det föredrag som hölls för Kungliga krigsvetenskapsakademien, avdelning IV, den 28 januari 2015.

Autonomi, är en central men svår term att definiera, eftersom den används på olika sätt i olika kontexter. Termen appliceras på allt från exempelvis dammsugare och gräsklippare till både enkla och mer avancerade robotar, samtidigt som man i de filosofiska seminarierummen diskuterar huruvida ens *människor* är autonoma. Vi har alla en intuitiv, vardaglig förståelse av termen, men kriterierna för autonomi är olika för gräsklippare och människor, där självständighet i någon mening ingår. Denna självständighet sträcker sig dock från ”kan fungera utan att fjärrstyras” till ”måste ha en fri vilja”, med allt vad det innebär. Autonomi när det gäller militära robotar verkar bestämmas utifrån relationen mellan människa och robot och huruvida människan är ”inne” i den så kallade beslutsloopen”, ”på” loopen eller utanför loopen, och på vilket sätt.[1] Intressant här är att notera att människan ibland kliver ur loopen frivilligt, eftersom hon helt enkelt är för långsam jämfört med många system, eller litar mer på systemet än sitt eget omdöme. När det gäller robotar och framför allt militära robotar, är det viktigt att reda ut och vara noga med autonomibegreppet för att man ska kunna hantera ansvarsfrågan – något vi kommer att återkomma till. En korrekt definition av

autonomi, med nödvändiga och tillräckliga villkor, handlar samtidigt om att ge kriterier för ansvar.

Att definiera termen ”robot” är inte heller enkelt. Om man exempelvis skulle definiera robot som ”en dator med sensorer och aktuatorer som tillåter den att interagera med den externa världen”, så skulle en dator kopplad till en skrivare anses vara en robot, vilket tycks kontraintuitivt.[2] Något mer krävs, som ”tänkande” eller ”agerande” i någon mening, som skiljer robotar från det automatiska, som exempelvis termostater eller den nämnda datorn kopplad till en skrivare.

Robotforskaren Ronald Arkin skiljer på *robot* respektive *autonom robot*. *Robot* är enligt Arkin en automatisk maskin som är kapabel till självständig perception, resonerande och agerande. En *autonom* robot är ”en robot som inte kräver direkt mänsklig inblandning utom för ’high-level mission tasking’ – en sådan robot kan fatta sina egna beslut konsistent med sitt uppdrag, utan att kräva direkt mänsklig auktorisering, även när de gäller beslut angående användandet av dödligt våld”.[3] Skillnaden verkar hänga på auktorisering snarare än autonomi, eftersom även enkla robotar kan förnimma, resonera och agera. Någon form av resonerande eller *tänkande*, krävs också.

Syftet med denna text är dock inte att ge några vattentäta definitioner av autonomi och robot, utan som redan nämnts peka på och belysa några av de filosofiskt intressanta frågorna och problemen kring militära robotar, där den här typen av definitioner är en del. Detta kommer stundtals att föra oss bort från konkreta militära robotar och in på filosofins område. Ett rimligt kriterium för autonomi är att man kan tänka i någon mening, och robotars förmåga att göra detta ska vi se närmare på i följande avsnitt.

Kan en maskin tänka?

Filosofiska resonemang om tänkande, och huruvida maskiner kan tänka, representerar en av de stora frågorna inom medvetandefilosofin. Man kan säga att det finns en skiljelinje mellan huruvida det är tillräckligt att en maskin beter sig *som om* den tänker, eller om man ska kräva något mer, något ”äkta” bakom ”kulissen”. Enligt Alan Turing kan man genom det så kallade Turingtestet, som beskrevs i hans numera klassiska artikel från 1950 – *Computing machinery and intelligence* – avgöra om en maskin kan tänka i relevant mening. Turingtestet går ut på att en människa ställer frågor till en människa respektive en dator som sitter bakom en skärm. Om utfrågaren inte klarar av att avgöra vem som är människa och vem som är dator, ska datorn anses ha klarat testet och kunna tänka – eller visa intelligens – i relevant mening. Turings idé är att man bör se exempelvis tankar utifrån vad de har för *funktion* i ett visst system. Om något – i det här fallet interna tillstånd av något slag – verkar fungera på samma sätt som (mänskliga) tankar, så ska dessa interna tillstånd anses vara tankar. Det spelar ingen roll hur de är konstruerade eller vad de består av.

Om vi har ett icke-mänskligt system – en robot – kanske vi spontant anser att det i en sådan inte kan finnas några tankar. Men Turing menar alltså att om det finns något som fungerar på samma sätt som exempelvis mänskliga tankar – att detta något t ex tar in indata och bearbetar dessa så att det kommer ut utdata på (utifrån sett) samma sätt som det skulle ske hos en människa, då fungerar detta interna tillstånd som tankar, och kan anses *vara* tankar. Turing anser med andra ord att man bör identifiera tankar med ett systems tillstånd definierat enbart utifrån vilken roll dessa har för att

producera ytterligare interna tillstånd och verbal output.

Detta sätt att se på interna tillstånd kallas ibland *funktionalism*. Det innebär att man alltså utgår från vilken funktion tillståndet har, snarare än hur det verkligen är beskaffat. Om något verkar fungera som en mänsklig tanke, så kan man anse det vara en tanke.

Finns det då någon dator som har klarat Turings test? Den 9 juni 2014 skrev *The Guardian* om vad man kallade det första lyckade Turingtestet i artikeln *Computer simulating 13-year old boy becomes the first to pass Turing test*. Fem maskiner testades på Royal Society i London i textbaserade konversationer, och ”Eugene Goostman” lurade trettiotre procent av utfrågarna att tro att den var en människa. Detta sågs som en milstolpe när det gäller artificiell intelligens (AI) – vissa anser att det inom AI inte finns något mer ikoniskt eller någon mer kontroversiell milstolpe än Turingtestet.[4] Andra menar emellertid att Turing inte gav tillräckligt specifika detaljer om testet, vilket innebär att det ovanstående lyckade testet i viss mån kan anses vara godtyckligt.[5] Det handlar t ex om att testet saknar specifikation om hur länge det ska pågå, hur sofistikerat det ska vara och liknande.

Det finns alltså vissa praktiska problem med Turingtestet. Men även om man skulle ta sig förbi dessa, kvarstår frågan huruvida det är viktigast hur ett tillstånd verkligen är beskaffat eller om det är viktigast vilken funktion tillståndet har i ett system. Även om man alltså ställer upp ett Turingtest som verkar vettigt och en dator klarar testet och därmed kan anses tänka i relevant mening, menar många att det ändå är något som skaver – att det ändå skulle vara något som saknas.

Istället kan man tänka sig en dator placerad i ett rum. Datorn är försedd med en översättningsmanual. En person som kan kinesiska stoppar lappar med frågor på kinesiska genom en lucka in i rummet. Datorn i rummet bearbetar dessa frågor med hjälp av manualen och skickar sedan ut svaren, även de på kinesiska, genom en annan lucka. Frågeställaren skulle antagligen tro att den – eller det – som sitter därinne i rummet och producerar svar, förstår kinesiska. John Searle, som är detta tankeexperiments upphovsman, menar dock att maskinen *egentligen* inte förstår. Det handlar inte om någon *genuin* förståelse, inte på samma sätt som om en människa som kan kinesiska hade suttit därinne och svarat på frågor. Maskinen använder ju bara en manual, där indata med hjälp av manualen omvandlas till utdata. Om det finns en förståelse är den i så fall vad man kallar *syntaktisk* – mekanisk, skulle man kunna säga – snarare än *semantisk*. Semantisk förståelse är den typ av förståelse som vi människor anses ha. Det är en sorts djupare förståelse, som innebär att man förstår *meningen* i något. Det är inte bara ettor och nollor, indata och utdata.

Enligt Searle kan maskiner inte ha verklig förståelse och han skilde mellan stark och svag artificiell intelligens (AI), där stark AI innebär att en maskin *verkligen* kan tänka och ha ett medvetande, medan svag AI bara kan bete sig *som om* den kan tänka eller ha ett medvetande, och frågan är vad vi bör kräva av robotar. Ska det räcka med att den kan bete sig som om den kan tänka?

Medvetande

När det gäller robotar och ansvar, så nämns ofta bristen på medvetande som ett viktigt kriterium. Robotarna kanske är intelligenta – i någon mening – och kan agera rationellt, men de har ingen uppfattning av att vara ett *jag* i betydelsen att de har några tankar om sig själva. Det finns dock inom

medvetandefilosofin olika synsätt på vad det innebär att ha ett medvetande.

Descartes menade att ett medvetande var lika med en odödlig själ som var av en annan substans än kroppen. Andra synsätt är att se medvetandet som *beteende* – vilket bland annat behaviorismen gör. Man kan också anse att medvetandet är detsamma som hjärnan – d v s att sätta likhetstecken mellan medvetandet och hjärnan, en teori inom medvetandefilosofin som kallas fysikalism. Eller så kan man se medvetandet som en dator. Det kommer in indata som behandlas, och ut kommer utdata.

Ett argument för Turings synsätt är det så kallade problemet med andra medvetanden.

Även om vi alla har samma upplevelse av att ha ett medvetande – ett jag – så kan man ställa sig frågan hur vi kan vara säkra på att andra människor upplever denna jagkänsla på samma sätt som vi gör. Det vi i själva verket gör är att dra slutsatsen att eftersom *vi* upplever oss ha eller vara ett jag, så gör nog andra det också – det är en slutsats som grundas på hur andra människor beter sig. När en annan person säger sig vara sugen på choklad och går och köper en chokladkaka, så tillskriver vi denna person mentala tillstånd liknande våra egna. Vi tänker att personen nog kände ett sug efter choklad som liknar vårt, och att processen för att tänka ut en plan för att få tag på choklad också liknade vår.

Man kan då fråga sig varför vi skulle göra på något annat sätt när det handlar om robotar, och om det inte är ”partiskt” (*biased*) att kräva att man måste vara *organisk* för att ha mentala tillstånd. Frågan är ju varför det är så självklart att något *artificiellt* inte skulle kunna ha ett medvetande. Vi vet ju faktiskt inte hur det känns för Lisa att vara Lisa eller för Lasse att vara Lasse – eller som Thomas Nagel poängterade i sin klassiska artikel *What is it like to be a bat?*,[6] att oavsett hur mycket tredjehandsinformation vi än har kan vi aldrig föreställa oss hur det är för en *fladdermus* att vara en fladdermus. Vi kan möjligen föreställa oss hur det är för en *människa* att vara fladdermus, men inte hur det är för en fladdermus att vara fladdermus. På samma sätt kan vi heller inte föreställa oss hur det känns för den avancerade Stjärnornas krigroboten C-3PO att vara C-3PO. Hur kan vi ens vara säkra på att den avancerade roboten faktiskt *inte* har ett medvetande? Och varför måste detta medvetande i så fall härstamma från något organiskt?

En annan fråga, som är nära besläktad med frågan om medvetande, är den om fri vilja, som i sin tur är kopplad till autonomi och ansvar. Kan en robot vara ansvarig för sina handlingar, eller är detta till och med en ”icke-fråga”, eftersom den alltid är programmerad av en människa och därmed människan därmed alltid är ansvarig för vad roboten gör? Detta är förmodligen extra viktigt att reda ut när det gäller just militära robotar.

Även om vi i framtiden skulle ha goda skäl att hålla en robot ansvarig for det den har gjort, kan man hävda att det inte vore någon *mening* med att göra det eftersom belöning eller straff inte skulle ha någon betydelse för en robot. Anledningen till att vi håller människor ansvariga är delvis för att påverka deçüras framtida beteende. Om vi får ett barn att inse att det har gjort fel – att det exempelvis har sårat någon och barnet kanske upplever skam- eller skuldkänslor, är chanserna goda att barnet beter sig annorlunda när det i framtiden befinner sig i en liknande situation. Här spelar alltså *känslor* av skam och skuld en viktig korrigerande roll.

Gordana Dodig Crnkovic och Baran Çürüklü är två forskare som inte anser att

det här med robotars ansvar är en ickefråga som omedelbart kan avvisas.[7] De hävdar att etiska aspekter runt artificiell intelligens inte har utforskats tillräckligt – delvis på grund av missuppfattningen att robotar bara gör vad de är programmerade att göra. Dodig-Crnkovic och Çürüklü anser att det är fel att säga att det vore *meningslöst* att hålla en robot ansvarig. Genom att titta på och ta intryck från säkerhetskulturen, vars främsta intresse är att lära sig från erfarenheter snarare än att skuldbelägga, kan man finna motargument mot detta. Att tillskriva robotar ansvar skulle kunna vara möjligt och till och med lämpligt för robotar som är självlärande.

Fri vilja och programmering

För att återgå till den fria viljan, som är ett rimligt kriterium för den grad av autonomi som krävs för ansvar, hävdar vissa att den fria vilja vi upplever att vi har i själva verket är en illusion. Om det skulle vara meningslöst att straffa eller klandra en robot som beter sig illa eller oetiskt, eftersom det bara är att stänga av den, kan man konstatera att de länder som använder sig av dödsstraff krasst uttryckt faktiskt ”stänger av” människor som begått tillräckligt allvarliga brott. Detta kan man invända mot inte bara på humanitära grunder, utan även enligt vissa sätt att se på den fria viljan. Vissa filosofer går så långt som att hävda att människor inte kan hållas moraliskt ansvariga för sina handlingar – och då vore det ju helt orimligt att någonsin hålla en robot ansvarig för sina handlingar.

Idén går ut på att våra handlingar är konsekvenser, d v s att de är styrda av naturlagarna och andra händelser i ett avlägset förflutet, som vi inte har någon som helst kontroll över. Vi kan ju t ex inte hållas ansvariga för vilka föräldrar vi har, och vilka föräldrar dessa i sin tur har – d v s vilka gener vi har, vilken uppfostran vi fick, vilken miljö vi växte upp i. Och vi kan sannerligen inte hållas ansvariga för hur naturlagarna ser ut. Den här synen kräver att man anser att världen är determinerad – att *determinismen* är sann. Det betyder att för varje händelse, även mänskliga handlingar, så finns det förutsättningar som inte kunde orsaka något annat än det som faktiskt hände. Även om vi *tror* att vi väljer mellan moroten och chokladkakan och väger argument för och emot och verkligen *kämpar* med vår längtan efter chokladkakan så kanske det i själva verket var så att alla dessa tankar, och känslan av att debattera med sig själv – mitt i ett val, som leder till ett beslut – bara är en illusion. I själva verket är det i någon mening förutbestämt vad vi kommer att välja, även om vi inte upplever att det är så. De som resonerar på det har sättet kallas ibland *inkompatibilister*. Det som är inkompatibelt är *fri vilja* och *determinism*.

Tanken är alltså att eftersom det inte är vårt ansvar vad som hände innan vi var födda, vilka gener vi har, hur vi har uppfostrats eller hur naturlagarna ser ut – så kan vi inte hållas moraliskt ansvariga för våra handlingar. Andra filosofer håller inte med om det här. De menar att fri vilja och determinism visst är kompatibla, och kallas därför *kompatibilister*. Enligt kompatibilisterna handlar frihet om att vara fri från vissa typer av hinder, så som att man exempelvis inte är fysiskt eller psykiskt tvingad att agera på ett visst sätt. Även om ens personlighet, preferenser och motivation bestäms av händelser man inte kan hållas ansvarig för, så menar kompatibilisterna att detta inte är nödvändiga krav för frihet i den relevanta meningen. Kompatibilistisk frihet handlar i stora drag om att välja och agera på det sätt man tror är bäst, *givet den man är* – att man kunde ha handlat annorlunda om man

hade haft en tillräckligt stark önskan, eller om man hade haft andra föreställningar om det bästa sättet att nå sitt mål. Man kan nog säga att det är den kompatibilistiska synen som gäller i samhället generellt.

Vi håller människor ansvariga om de verkar vara vid sina sinnens fulla bruk – om de verkar kunna resonera enligt åtminstone basal logik – och inte har blivit uppenbart tvingade. Det går ju knappast att skylla på sin olyckliga barndom i en rättgång. Frågan är då om robotar kan involveras i en sådan kompatibilistisk syn på fri vilja. Om det handlar om att en robot kunde ha agerat annorlunda om den hade varit programmerad annorlunda så verkar det finnas en avgörande skillnad mellan människor och robotar. Människan är ju inte programmerad – åtminstone inte bokstavligen. Men en avancerad robot skulle lika gärna kunnat ha andra föreställningar eller önskningar. Om den är programmerad att skydda den egna armén eller det fartyg den befinner sig på, men har frihet att uppnå detta övergripande mål på vilket sätt den vill, kanske den inte är mer programmerad än människan som är biologiskt programmerad att skydda sig själv och sin familj. Det finns dock vissa som menar att resonemang om ”bokstavlig” och ”biologisk” programmering inte räcker för att inkorporera robotar i den grupp som är tillräckligt fria för att kunna hållas ansvariga för sina handlingar – att robotar saknar något fundamentalt, som är högst relevant för ansvar. Även om utvecklingen går mot ökad autonomi, och robotar kanske kan bli så avancerade att de har en fri vilja på *ungefär* samma sätt som människor anses ha i vårt samhälle, så är det ändå något som saknas. Robotar saknar känslor, de saknar det organiska. En robot kan ju t ex aldrig få självmedvetande eller en egen vilja. Eller kan den? I och med denna fråga är vi tillbaka i samma medvetandefilosofiska frågeställning som tidigare. Syftet här är dock inte att bestämma huruvida robotar bör hållas ansvariga eller inte, utan att visa på den filosofiska problematik som omger robotar, autonomi och ansvar.

Etiska frågor

De etiska frågorna rörande militära robotar kan delas in i olika kategorier. En första kategori gäller etiska aspekter på användandet av robotar i krig. Dit hör frågor om hur olika sorters robotar, med olika grader av autonomi – och med utgångspunkt i olika moraliska normativa teorier – kan påverka tolkningen av krigets lagar, samt huruvida autonoma robotar påverkar rättfärdigandet att döda, som är implicit i krigets lagar – och huruvida dessa bör korrigeras i och med det ökade användandet av alltmer autonoma robotar. En annan kategori handlar om hur militära robotar med en hög grad av autonomi ska programmeras för att bete sig ”etiskt”.

Etiska frågor rörande användandet av robotar

När man ställer sig frågan huruvida det är etiskt att använda exempelvis UAV:er finns det olika utgångspunkter, som exempelvis realism (enligt vilken man inte ska applicera moraliska termer som rättvisa eller rättfärdigt på krig, eller staters förehavanden internationellt. Moral är en lyx en stat inte kan hålla sig med), pacifism (enligt vilken moraliska koncept kan appliceras på krig, och frågan huruvida ett krig är rättfärdigt är en meningsfull fråga. Svaret är dock att man aldrig ska kriga, åtminstone inte enligt absolut pacifism), och teorin om rättfärdigt krig, *just war theory*. Det är den sistnämnda som ligger till grund för krigets lagar, som är kodifierade i Genève- och Haagkonventionerna och som framför

allt är intressant när det gäller militära robotar.

I just war theory skiljer man mellan regler för att starta krig (*jus ad bellum*), som t ex att kriget måste vara rättfärdigt (*just*), och regler inom kriget (*jus in bello*), som t ex att en attack måste vara nödvändig och inte får vara överdriven, samt vilka vapen som inte får användas (exempelvis biologiska). Ett argument mot att använda exempelvis UAV:er är att tröskeln för att starta krig skulle sänkas eftersom man riskerar färre av de egna soldaternas liv, och kriget riskerar att bli mer som ett dataspel. Man kan också fråga sig om vissa attacker verkligen skulle ha utförts om motståndaren hade kunnat använda UAV:er. Ett annat problem är att det kan uppstå känslor av orättvisa, som i sin tur kan leda till hämndaktioner, om användandet av UAV:er upplevs som fegt. En annan fråga rör hur man i krigets lagar ska hantera det faktum att operatörer sitter och styr UAV:er från andra sidan jorden. Skulle det vara tillåtet att attackera en UAV-operatör som är på väg till eller från sin arbetsplats? Det kan också uppstå hemliga ”krig”, t ex när CIA använder UAV:er – och då har man eventuellt gått ifrån exempelvis regler som att krig ska föras mellan stater, som är ett av kriterierna i *jus ad bellum*.

Argument kopplade till *jus in bello* handlar om avtrubbning och tekniska fel. Det har dock visats att UAV-operatörer lider av posttraumatisk stress, i vissa fall i högre grad än soldater som är på plats. Detta förklaras delvis av att operatören kanske följer sitt mål i dagar eller veckor och ”lär känna” personen.

Etisk tolkning av krigets lagar

I krigets lagar finns några termer som är ganska vaga och som kan tolkas olika beroende på vilken normativ moralisk teori man grundar analysen på – och robotar kan i sin tur påverka just detta. De termer som åsyftas är framför allt rättfärdigt (*just*) i *jus ad bellum*, och nödvändigt (*necessary*) eller överdriven (*excessive*) i *jus in bello*.

Även om det är ganska utförligt specificerat vad som är ett rättfärdigt skäl att starta ett krig, så är denna regel öppen för tolkning. Rättfärdigt (*just*) är inte detsamma som moraliskt riktigt (*right*), men de hör ihop. Enligt en utilitaristisk tolkning av ett rättfärdigt skäl kan rättfärdigt kopplas till maximerad nytta, och det är möjligt att argumentera för en lokal snarare än global nytta. Land A kanske lider på grund av några faktorer som gör det rättfärdigt att starta krig mot land B, och A säger att om de attackerar B, så kommer B att lida, men om B ger upp snabbt och ger A en del av sitt territorium, så blir den totala summan av lidandet mindre än A:s lidande idag. Ett exempel på en *pliktetisk* tolkning av ett rättfärdigt skäl att starta ett krig skulle kunna vara uppfattningen att man måste försäkra sig om att alla människor lever under en särskild religions lagar eller kanske Kants kategoriska imperativ, d v s att bara agera enligt en maxim som man vill se upphöjd till universell lag. Det skulle mycket väl kunna ge en tolkning som gör det rättfärdigt av ett land att starta ett krig.

Även om alla regler i *jus ad bellum* måste uppfyllas, så skulle det land som har UAV:er kunna argumentera för att dessa faktiskt ökar chansen att uppfylla exempelvis villkoret ”rimlig chans att lyckas”. En *dygdetisk* tolkning av ett rättfärdigt skäl att starta ett krig är svårare att slå fast. Att agera rättfärdigt skulle förstås vara dygdigt, men dygdetiken har ju, som vi sett, problem med att ge någon klar handlingsvägledning. Det här indikerar att termen rättfärdigt kan ”kapas” av ett land med drönare, d v s att man hittar en lämplig tolkning av ett rättfärdigt

skal att starta ett krig. Termen ”*nödvändig*” är eventuellt ännu vagare än ”*rättfärdig*”. Vad som är nödvändigt kan, enligt utilitarismen, betyda i stort sett vad som helst, eftersom denna teori ofta kopplas samman med talesättet ”målet helgar medlen”. Det finns inga handlingar som är kategoriskt förbjudna enligt utilitarismen. Det är dock viktigt att poängtera att utilitarismen, strikt eller ej, sällan appliceras i krig. För det första är det förstås inte så att ett land anser att fiendens liv är varda lika mycket som de egna liven. För det andra är det inte troligt att stridande parter skulle göra precis vad som helst för att vinna. Det kan emellertid vara så, om man frågar i en specifik situation – exempelvis om det är nödvändigt att döda så många – att en utilitarist säger ja om den totala summan av nytta är högre än om de inte hade dödats. Förespråkare för pliktetik kan resonera på ett liknande sätt genom att använda *dubbla effekt-doktrinen*, enligt vilken det inte är tillåtet att skada någon med avsikt, men att det kan vara moraliskt försvarbart att genomföra handlingar som man har goda avsikter med, och där skadan är något man förutser men inte avser. Man kan t ex resonera som så att civila offer är tillåtna, men att det är förbjudet att sikta på civila – att avsiktligt skada dem. Men om de råkar stå i vägen, kan det vara tillåtet.

Dygdetik kan, när det gäller att tolka vad som är nödvändigt, framstå som ett skydd mot grymheter när de stridande är förblindade av önskan att segra. Det är dock inte lätt att få fram ett säkert svar på vad dygdetiken anser vara rätt. Rättvisa (*fairness*) är förmodligen en del av ekvationen, vilket gör det problematiskt om bara den ena sidan skulle drabbas av dödsoffer. Enligt äldre versioner av dygdetik, som härstammar från Aristoteles, är många av dygderna kopplade till krigaren. De gamla kardinaldygderna rättrådighet, tapperhet, vishet och måttfullhet kan ju i ganska hög grad kopplas till ett krigarideal, särskilt om man lägger till mod, som är en av de klassiska religiösa dygderna. Det kan vara intressant att jämföra dessa med dagens ideal för officerare, som de ser ut i ledstjärnorna for svensk officersetik.[8] Dessa regler påminner om Aristoteles resonemang om dygder, och markerar också en skillnad mellan olika dygdetiker.

Där exempelvis Aristoteles talar om dygder i första hand har dygdetikern Rosalind Hursthouse gjort försök att ge ett riktighetskriterium för dygdetiken, som inte ser till specifika dygder eller listor av dygder som grund för vad som är moraliskt rätt (även om hon indirekt gör detta i talet om den dygdige).[9] Dygdetik kan alltså ge olika svar på frågan om vad som är nödvändigt. Det är också möjligt att hävda att det är nödvändigt med exempelvis en drönarattack, eftersom man vinner kriget snabbare och det blir färre döda om bara den ena sidans soldater dör (istället för båda).

Överdrivet (*excessive*) är nära kopplat till *nödvändigt* och kan fungera som bromskloss för den som är för fokuserad på målet för att bry sig om medlen. Här kan ju en utilitaristisk, pliktetisk eller dygdetisk tolkning skilja sig åt. Det finns dock några kriterier. Man får t ex inte göra något som inte direkt har med krigets seger att göra, eller utföra en handling som har liten betydelse för slutmålet, i jämförelse med den skada man kommer att åsamka. Walzer tolkar detta som att det finns två kriterier: dels det kriterium som handlar om segern i sig, eller vad som brukar kallas militär nödvändighet. Det andra kriteriet har med proportionalitet att göra, och att man där ska ta hänsyn inte bara till enskild skada utan till mänskligheten i stort. Även här kan en utilitarist hävda att det är bättre att använda UAV:er eftersom då bara den ena sidans soldater dör

istället för båda. Ett land som har UAV:er kan göra en utilitaristisk tolkning, ett land utan drönare kan istället göra en pliktetisk eller dygdetisk tolkning och komma fram till olika slutsatser. Dygdetiker kan mena att UAV:er är ett ”fegt” vapen och att det därför bör förbjudas, ungefär som kärnvapen eller biologiska vapen är förbjudna.

Resonemang gällande det oetiska och fega i ökat avstånd har som sagt förts ända sedan man började använda pil och båge. Begreppet överdrivet kan i viss mån kopplas till Aristoteles tankar om den gyllene medelvägen och även till den klassiska kardinaldygden måttfullhet. Den amerikanske moralfilosofen R.B. Brandt har föreslagit att man ska använda Rawls' *okunnighetsslöja* om man ska bestämma vilken bas som är lämplig för att tolka krigets lagar.[10] Han menar att bakom slöjan – där man inte vet i vilket land man kommer att födas, om man blir rik eller fattig, föds i ett rikt, högteknologiskt land eller ett utvecklingsland och så vidare – skulle de flesta välja utilitarismen som bas. Han tror att folk skulle välja de regler som skulle maximera långsiktig förväntad nytta för nationer i krig. Men man kan fråga sig vilken nytta det är som ska tas hänsyn till – den totala summan hos båda länderna, eller hela världen? Och i så fall, med vilket tidsperspektiv? Man hamnar snabbt i klassiska utilitaristiska frågeställningar. Kanske är Kants kategoriska imperativ bättre bakom okunnighetsslöjan, eftersom man inte vet om man kommer att bo i ett land med robotar, hur avancerade dessa robotar kommer att vara, eller inte.

Ett annat problem med Brandts påstående, och ökningen av robotar, är att utilitaristerna får svårt att inkorporera nationaliteter i kalkylen. En kantiansk tolkning kan dock tillåta robotar genom doktrinen om dubbel effekt. En intressant fråga när det gäller tolkningen av krigets lagar är huruvida man på något sätt borde specificera vilken normativ teori dessa lagar bör baseras på, eller om man bör göra reglerna tydligare. Walzer diskuterar detta. Han menar att utilitarismen saknar kreativ kraft när det gäller krig – att den bara kan konfirmera existerande seder och konventioner, vilka de nu är, eller föreslå att man ska förbise dessa. Walzer menar att rättighetsetik är att föredra i det här läget, men då måste man redogöra för hur rättigheter fungerar i krig och strid – hur de erhålls, förloras, återfås och byts. Detta försöker Walzer göra i *Just and Unjust Wars*.[11] De grundläggande rättigheter han utgår från är bland annat mänskliga rättigheter om liv och frihet. Han säger bl a att ”the ban on rape and murder is a matter of right”.

En annan intressant fråga gällande robotar och krigets lagar är rättfärdigandet att döda, och om det etiska rättfärdigandet att döda i krig blir påverkat eller undermerat om människor ersättas av robotar i hög utsträckning. Just war theory, som krigets lagar är ju byggda på, har ju en strikt uppdelning mellan *jus ad bellum* och *jus in bello* – där alla kombattanter anses vara moraliskt jämlika när kriget väl har kommit igång – *the moral equality of combattants* (MEC). En teori är att MEC och i förlängningen rättfärdigandet att döda bygger på ett ömsesidigt utsättande för risk – att man riskerar det man hotar, och om ena sidan slåss enbart med robotar kan man tänka sig att denna jämvikt hotas.

Den etiska roboten

När det gäller att få en robot att bete sig etiskt, kan man tänka sig olika tillvägagångssätt. Dels kan man använda sig av en så kallad top-down-metod, där man skulle kunna använda en normativ moralisk teori – varav några redan har nämnts i denna text, som utilitarism, pliktetik eller dygdetik. Ett

mer känt exempel på regler som programmeras in i robotar är Asimovs lagar (1. En robot får aldrig skada en människa eller, genom att inte ingripa, tillåta att en människa kommer till skada, 2. En robot måste lyda order från en människa, förutom om sådana order kommer i konflikt med första lagen, 3. En robot måste skydda sin egen existens, såvida detta inte kommer i konflikt med första eller andra lagen). Men ett problem med dessa är att de inte är tillräckliga för mer avancerade robotar, som ställs för mer komplicerade och varierande situationer och dilemman.

Man kan t ex tänka sig det så kallade trolley-exemplet – ett klassiskt tankeexperiment som ofta tas upp på grundkurser i filosofi. Det går ut på att man tänker sig att man står vid en växelspak och kan styra in ett herrelöst, skenande tåg som är på väg att köra över fem personer längre fram – som man på intet sätt kan varna eller kommunicera med – eller in på ett annat spår, där det istället står en person. Om man drar i spaken medverkar man alltså till att en dör istället för fem. Frågan är om man skulle dra i spaken eller inte, och i så fall av vilka skäl. Därefter ställs man inför en liknande situation, men där man istället för att dra i en spak ska knuffa en storväxt person framför tåget och på så sätt få det att stanna. De flesta brukar svara att de drar i spaken, men inte knuffar mannen, och får då ofta problem med att förklara varför.

De flesta skulle nog föredra att ett autonomt tåg skulle välja att växla in sig själv på det spår där bara en person kommer att dö. Man kan dock fråga sig om det är rätt att roboten alltid ska se till att flest överlever, eller om vi ska låta den ta hänsyn till vissa aspekter. Är det t ex bättre att låta en cancerforskare leva än fem dömda mördare som har lyckats rymma från ett fängelse? Där ens ”värde” för samhället scannas in, och används av en robot som ska fatta beslut? Vill vi att en robot ska ta hänsyn till sådant, eller bara räkna antalet liv? Ska alla människor alltid vägas lika? Det här är typiska exempel på etiska frågor som inte har några självklara svar, eller som åtminstone kräver att man reflekterar innan man fäller sitt omdöme – och ger goda, genomtänka argument för detta.

En annan metod är bottom-up-metoden – där robotar lär sig moral ungefär som ett barn. Syntetiska känslor, exempelvis skuldkänslor som styr framtida beteende, kan då vara ett sätt att åstadkomma detta.

Man kan också fråga sig vem som ska bestämma vilken moralisk teori som ska implementeras om man bestämmer sig för att använda top-down-metoden. Bör man ha ”computational complexity”, d v s hur svårt det är att omvandla en viss teori till algoritmer, när man bedömer normativa teorier för artificiell moral, och förkasta vissa teorier för att de blir för komplexa matematiskt? Vilka kan de långsiktiga konsekvenserna bli av ett sådant förfarande? Och i vilken utsträckning kan ”computational tractability” (hanterbarhet) vara ett hinder för samhällets etiska orientering? Hur medier väljer att framställa olika företeelser kan vara en faktor när det gäller vilken normativ teori som implementeras. Man kan också tänka sig en databas där människor kollektivt bestämmer vilken handling som är önskvärd i ett stort antal situationer – ”demokratisk etik”.

Robotarnas framfart gör att vi tvingas besvara sådana filosofiska frågor, och det är viktigt att reda ut sådana här frågor innan man bestämmer sig för att helt förbjuda utvecklandet av autonoma robotar i den demokratiska västvärlden.

Författaren är fil dr och tjänstgör vid Försvarshögskolan.

Noter

1. Systemet Aegis har exempelvis olika lägen, som semiautomatisk, automatisk special, automatisk, samt *casualty*.
2. Lin, Patrick; Abney, Keith och Bekey, George: *Robot Ethics – The Ethical and Social Implications of Robotics*, MIT Press, Cambridge, Massachusetts 2012, s 17.
3. Arkin, Ronald C: *Governing Lethal Behavior in Autonomous systems*, CRC Press, 2009, s 50-51.
4. Exempelvis professor Kevin Warwick från University of Reading.
5. Kurzweil har föreslagit regler för detta: "A wager on the Turing test rules": *www.kurzweilai.net/a-wager-on-the-turing-test-the-rules*, (2015-01-30).
6. Nagel, Thomas: *The Philosophical Review*, vol 83, nr 4, 1974, s 435-450.
7. Dodig-Crnkovic, Gordana och Çürüklü, Baran: "Robots – Ethical by Design", *Ethics and Information Technology* , vol 14, nr 1, 2011, s 61-71.
8. (Min koppling till äldre dygder inom parentes): 1. När du valde att bli officer tog du på dig att värna om vårt lands militära säkerhet – i strid om nödvändigt. Som officer är du ledare och utövar makt – såväl i fred som i krig – ett förtroende du fått av svenska folket. (*Rättrådighet*) 2. I fred skall du förbereda dig själv och ditt förband för uppgifter under krig, fred och kris. […] Få yrken innefattar en så tung ansvarsbörda. (*Tapperhet)* 3. Väpnade konflikter är ett ont. Men anfalls vi, och måste tillgripa vapen for att värja oss, är du medansvarig för att målet nås med minsta möjliga skada. Tank i tid på dem som skall få bära konsekvenserna av dina handlingar (*Tapperhet, rättrådighet och måttfullhet*). 4. Situationer kan uppstå då varken order eller regler ger dig full vägledning. Du möter moraliska dilemman, där ingen kan tala om för dig hur du skall handla. Då behöver du en genomtänkt etik. (*Vishet*) 5. En försvarsmakt måste bygga på att order lyds. Men orderlydnad kan aldrig rättfärdiga handlingar som strider mot etik och folkrätt. […] (*Vishet och rättrådighet)* 6. Lev upp till de höga krav som din roll medför. Ställ aldrig högre krav på dina medarbetare än på dig själv. […] bevara det svenska folkets förtroende och respekt […]. (*Rättrådighet*) 7. Din yrkesetik ska vara förankrad i det idéarv som bär upp vårt samhälle. Humanism och demokrati är centrala beståndsdelar i detta arv. Förakt för oliktänkande eller för andra etniska grupper är lika oförenligt med din yrkesroll som obefogad användning av makt och våld. (*Vishet och måttfullhet).* Källa: Lundberg, Anders: *Vägar till svensk officersetik,* Försvarshögskolan, Stockholm 1997, s 14.
9. Hursthouse, Rosalind: *On Virtue Ethics*, Oxford, Oxford University Press, 2002.
10. Brandt, Richard Booker: "Utilitarianism and the rules of war", *Philosophy and Public Affairs*, vol 1, nr 2, 1972, s145-165.
11. Walzer, Michael: *Just and Unjust Wars*, The Perseus Books Group, New York 2006.