# När maskiner fattar beslut – vem är ansvarig?

Maria Hedlund

**Abstract**

Robots making decisions on social benefits, driverless cars causing traffic accidents, search engines presenting a selected narrow picture of the world – the rapid development of AI technology gives rise to machines that makes their own decisions, without direct influence from humans, but who is responsible for what a machine does? Can the machine itself be responsible? The aim of this article is to discuss and problematize responsibility relations when machines make decisions. The overarching question is whether machines can be responsible and if so, under which circumstances. Drawing on theories on responsibility, machine ethics, robot philosophy, and on recent AI development, the article demonstrates how functionalistic arguments can lead to the conclusion that machines are responsible for their actions, while approaches building on philosophical understandings of autonomy and agency rules out machine responsibility. Unless the machine is conscious, human actors always need to be responsible for decisions taking by machines. However, as self-improving systems increase machine autonomy and decrease human control, the question is raised whether we are witnessing an emerging responsibility gap, or if this development rather describes a situation of blurred responsibility, in which responsibility needs to be distributed between many different actors – AI developers, programmers, distributors, users, policy makers.

Artificiell intelligens (AI) har rötter i 1950-talets försök att utveckla datorprogram och robotar som kan efterlikna mänskliga förmågor som att använda språk, lösa problem och förbättra sig själva (Russell 2019; se även Kalle Åströms bidrag i denna volym) och har kommit att användas som ett samlingsnamn på teknik som med någon form av autonomi gör saker som människor gör, men ofta mycket snabbare. AI-teknik ligger bakom många tillämpningar som de flesta av oss möter mer eller mindre dagligen och snabbt har tagit för givna

Maria Hedlund är verksam vid Statsvetenskapliga institutionen, Lunds universitet.
E-post: maria.hedlund@svet.lu.se

såsom sökmotorer som vet var du befinner dig och minns vad du har sökt på tidigare, mobiltelefoner som låses upp med hjälp av ansiktsigenkänning, strömningstjänster som tipsar om musik eller teveprogram som du kan tänkas tycka om, eller sociala medier, som visar inlägg från användare med liknande åsikter. Andra exempel på områden där AI-teknik används är rekrytering, självkörande fordon, automatiserad börshandel och inom vården, där bildanalys används för att ställa diagnoser. AI har blivit något av ett *buzzword* i den offentliga debatten och man kan till och med stöta på uttrycket ”en AI” som i exempelvis ”En AI vi förstår” (Voister), ”Fejkade vd:ns röst med en AI – kom över 2,4 miljoner” (Techworld) och ”Kan en AI beräkna och förutsäga progressiv utveckling?” (Quora), vilket ger intrycket att AI i sig uppfattas som ett självständigt subjekt. Det är inte alldeles långsökt, eftersom AI, som antyds ovan, ofta förknippas med autonomi, dvs. självständigt agerande. Detta blir tydligt i den definition av AI som Europeiska kommissionen använder:

> Artificiell intelligens (AI) syftar på system som uppvisar intelligent beteende genom att analysera sin omgivning och agera – med någon grad av autonomi – för att uppnå specifika mål (EU-kommissionen 2019, förf.:s översättning).

Exakt hur självständigt ett AI-system behöver vara för att kallas autonomt är emellertid omstritt. Räcker det att maskinen, så snart människan ”tryckt på knappen”, utför en förprogrammerad rutin? Eller krävs att maskinen inte bara gör något på egen hand, utan också självständigt kan bestämma vad det är den ska göra och hur den ska göra det (jfr Burri 2016)? Maskiner som kan agera (någorlunda) självständigt väcker frågor om ansvar. Vem är till exempel ansvarig för beslut som en självlärande algoritm fattar? När Googles sökmotor lär sig vad du brukar söka efter och ger dig träffar som kan tänkas ligga i linje med dina preferenser och inte låter dig se andra relevanta träffar, kan det begränsa din bild av världen och i förlängningen din förståelse av andra människors perspektiv. Kan vi då säga att Googles algoritm är ansvarig för dessa sökträffar? När en självkörande bil orsakar en olycka på grund av att smuts på stoppskylten får bilen att uppfatta trafikmärket som en 90-skylt och ökar farten – kan vi då säga att bilen är ansvarig? Och vad skulle det i så fall innebära för bilen? Den skulle knappast lida av ett straff, än mindre förstå att det var ett straff.

Vanligtvis brukar vi säga att maskiner inte kan vara ansvariga för sina handlingar. Det har dock från politiskt håll rests krav på att robotar ska kunna hållas ansvariga för skador som de orsakar. I januari 2017 föreslog EU-parlamentets juridiska utskott att robotar skulle få status som ”elektroniska personer” för att kunna hållas ansvariga för sina handlingar (EP 2017: 59F). Detta förslag var dock inget som togs vidare av EU-kommissionen, som sedan hösten 2017 arbetar med att komma fram till ett europeiskt förhållningssätt till AI (EC

Communication 2018).[1] Det faktum att politiska aktörer lyfter frågan, kan dock ses som ett erkännande av att utvecklingen inom artificiell intelligens och robotik utmanar synen på ansvar (Gunkel 2017: 5).

För att kunna ta ansvar, krävs ett handlande subjekt och någon form av agentskap, eller agens, det vill säga någon form av kontroll över sina handlingar (Kane 2002: 698). När vi talar om juridiskt ansvar, är också subjektets intentionalitet avgörande (Lanteri 2009). Den (eller det) som saknar båda dessa egenskaper (agens och intentionalitet) kan därför inte hållas juridiskt ansvarig (HRW 2015: 18). Men även om en maskin inte kan hållas juridiskt ansvarig, kanske den kan ha moraliskt ansvar för vad den gör?

Denna artikel tar visserligen utgångspunkt i observerbara tekniska landvinningar (inom AI och robotik), men syftet är främst att på ett principiellt plan diskutera och problematisera ansvarsförhållanden när maskiner fattar beslut. Den övergripande frågan som ställs är om maskiner kan vara ansvariga för beslut som de fattar, och i så fall, under vilka omständigheter. Med maskin avses här såväl mjukvara (algoritmer) som hårdvara (robotar) som utför uppgifter med en viss grad av artificiell intelligens.

Framställningen tar sin början i ansvarsbegreppet – vad vi vanligtvis menar med ansvar och vilken sorts ansvar som är relevant här – för att därefter diskutera i vilken utsträckning beslutsfattande maskiner kan ses som ansvarstagande eller ansvarsbärande agenter. Den teknologiska utvecklingen har inneburit allt självständigare maskiner och allt mindre direkt mänsklig kontroll över vad maskinerna faktiskt gör och hur de kommer fram till vad de ska göra. Innebär detta att människors ansvar successivt övergår till maskiner? Vad skulle det i så fall betyda, rent konkret, att maskiner har ansvar? Har det överhuvudtaget någon betydelse för ansvarsfrågan vad maskiner kan göra? Aspekter som autonomi och agens blir här centralt, men också relationen människa – maskin.

## Vilken sorts ansvar?

När vi talar om ansvar tänker vi ofta på kausalt ansvar, det vill säga att den som har orsakat ett visst utfall också är ansvarig – svarar för – detta utfall. För att vara ansvarig, räcker det dock inte att ha orsakat utfallet. Om en aktör ska kunna hållas ansvarig krävs också att hon har eller kan förväntas ha adekvat kunskap om de kausala förhållandena och konsekvenserna av att handla eller att inte handla (Thompson 1987: 47). En förutsättning för ansvar är också att aktören har haft möjlighet att handla annorlunda (Thompson 1987: 40). För att tillskrivas ansvar måste aktören alltså kunna bindas till en handling vars

1 Ett annat exempel är Saudiarabien, som i oktober 2017 gav medborgarskap till roboten Sofia (Hart 2018). Det väckte en hel del kritik att ett av världens mest förtryckande länder gav en robot rättigheter som vida överstiger de rättigheter som landets kvinnor har, men det hela var sannolikt ett sätt att få uppmärksamhet och att symbolisera att Saudiarabiens ekonomiska framtid inte bara handlar om olja.

kausala förhållanden hon förstår (Adam & Groves 2011: 18) och som hon kan kontrollera (Fischer & Ravizza 1998: 14). Det går med andra ord inte att hålla någon ansvarig för en handling som vederbörande utför under tvång (Talbert 2008: 517).

En distinktion kan göras mellan kausalt ansvar och moraliskt ansvar, där skillnaden ligger i att moraliskt ansvar också involverar skuld och klander (Miller 2001: 458). Moraliskt ansvar förutsätter en moralisk agent, alltså en aktör som är medveten om – har kunskap om – en handlings moraliska natur (Held 1970: 97). Med moralisk avses här för enkelhetens skull normer om vad som är rätt och gott i en viss kontext, normer som kan, men inte behöver vara juridiskt bindande. En aktör som oavsiktligt orsakar en dålig situation kan således tillskrivas kausalt men inte moraliskt ansvar, eftersom hon inte har gjort något klandervärt (Talbert 2008: 516). Juridiskt ansvar skulle kunna ses som ett specialfall av moraliskt ansvar, eftersom det är en viss typ av normer som avgör om en handling är klandervärd eller inte, nämligen legala normer. Dessutom är legala normer något vi är skyldiga att känna till; vi befrias inte från juridiskt ansvar med hänvisning till att vi inte kände till lagarna.

Resonemanget så här långt utgår från att vi talar om huruvida ansvar kan läggas på en individ. Flera individer som agerar tillsammans komplicerar ansvarsfördelningen. Det kan ifrågasättas varför vi överhuvudtaget behöver bry oss om ansvarsfördelning mellan flera individer när vi funderar över om maskiner kan vara ansvariga. Frågan om maskiners ansvar kommer dock till viss del att behöva handla om deras ställföreträdares ansvar, vilket aktualiserar en diskussion om kollektivt ansvar. Ett kollektiv är något mer än ett aggregat av flera individer, då medlemmarna i ett kollektiv har någon form av band till varandra som aggregatet saknar (Gilbert 2006: 95, 102). Frågan är om kollektivet kan hållas ansvarigt som ett kollektiv, eller om ansvaret ska fördelas mellan de olika medlemmarna, och i så fall hur.

Ett omstritt sätt att hålla ett kollektiv ansvarigt är att tillskriva ansvar till kollektivet som en enhet (Thompson 1987: 44). Huvudfrågan här är huruvida ett kollektiv kan betraktas som en moralisk agent, vilket är en förutsättning för moraliskt (och juridiskt) ansvar. En organiserad grupp skulle möjligen kunna betraktas som en moralisk agent och därmed kunna tillskrivas ansvar på motsvarande sätt som en individ (Held 1970: 97), det vill säga på grundval av kausalitet, medvetenhet, och kontroll. En annan komplikation med kollektivt ansvar är huruvida och under vilka förutsättningar individer kan hållas ansvariga för andra individers handlingar (McGary 1986: 79). En sådan förutsättning är gruppsolidaritet, vilket, även om inte alla medlemmar är klandervärda, skulle utgöra en grund för ställföreträdande kollektivt ansvar (Feinberg 1970: 61–62, 73). I hierarkiska organisationer kan formella positioner avgöra vem som är (juridiskt) ansvarig för vad kollektivet gör (May and Hoffman 1991: 11). En arbetsgivare kan exempelvis vara ansvarig för vad de anställda gör som

om arbetsgivaren själv hade utfört handlingen. Ett sådant ställföreträdande ansvar motiveras av att arbetsgivaren kontrollerar vem som anställs (kontrollargument) och har större möjlighet att ta (ekonomiskt) ansvar (positionsargument) (Feinberg 1968: 675–676).

Det som hittills har sagts om ansvar avser bakåtblickande ansvar, som handlar om att identifiera den som är ansvarig för något dåligt som redan har hänt i syfte att på något sätt straffa den felande. I politikens värld syftar detta ofta på att den ansvariga ska avgå. Allmänna val sägs till exempel vara medborgarnas sätt att ”utkräva ansvar” av politikerna. Inom juridiken hittar vi bakåtblickandet i olika skäl till varför den som brutit mot lagen ska straffas: dels handlar straffet om ett sätt för brottslingen att sona sitt brott genom någon form av återbetalning (retribution) till samhället, dels syftar straffet till att ”kommunicera samhällets fördömande av det klandervärda beteendet” (Wahlberg 2019: 125). Rättfärdigande av straff genom att framhålla dess preventiva funktion är mer framåtblickande, både när det handlar om att förhindra den individuella brottslingen från att begå brott, i alla fall temporärt, och om att avskräcka andra från att begå brott (Wahlberg 2019: 125).

En framåtblickande syn på ansvar, såväl juridiskt som moraliskt, tar fasta på just vad som ska hända i framtiden. Distinktionen mellan å ena sidan att *bära* ansvar, med betydelsen att klandras för något som redan har hänt, och å andra sidan att *ta* ansvar, med betydelsen att göra saker nu med en blick på deras framtida konsekvenser (Rovane 2014: 12), är belysande. På engelska görs denna åtskillnad i språket, där *accountability* eller *liability* står för bakåtblickande ansvar, medan *responsibility* kan, men inte behöver, syfta på framåtblickande eller prospektivt ansvar. På svenska talar vi ibland explicit om ansvarsutkrävande när vi syftar på bakåtblickande ansvar, men ofta är den bakåtblickande betydelsen underförstådd när vi pratar om ansvar. Det blir därför viktigt att särskilt framhålla när framåtblickande ansvar åsyftas.

En poäng med framåtblickande ansvar är att det, istället för att invänta att dåliga saker händer, identifiera den som har gjort fel och straffa den felande, riktar fokus mot hur vi kan *undvika* att dåliga saker händer. Framåtblickande ansvar handlar alltså om att ta ansvar för att göra gott i framtiden, vilket innebär att kausalitet inte får samma framträdande roll i bedömningen av vem det är som är ansvarig. Kausalitet åberopas ibland som skäl för att peka ut vilka aktörer som bör ta ansvar för framtiden, exempelvis i debatten om kompensation för klimatförändringarnas effekter. Det som då åsyftas är att de som har orsakat dessa effekter också bör vara ansvariga för att kompensera dem som drabbas av dem. Kausala förhållanden är komplexa, men även om de skulle vara svåra eller i praktiken omöjliga att spåra (jfr Albertsen 2015), finns de. Svårigheter att identifiera de ansvariga aktörerna, som när det gäller klimatförändringar både är många – länder, regioner, företag, beslutsfattare – och inte alltid längre existerande, gör emellertid att kausalitet inte förefaller vara någon

framkomlig väg för att fördela detta ansvar. Dessutom kan man fundera över om detta överhuvudtaget handlar om framåtblickande ansvar, utan snarare är ett fall av retribution (jfr ovan och Wahlberg 2019).

Andra grunder än kausalitet förefaller rimligare för fördelning av framåtblickande ansvar. Makt och resurser kan utgöra en sådan grund. En aktör som befinner sig i en position som gör att hon har möjlighet att agera eller som besitter kapacitet att agera kan ha en moralisk skyldighet att förbättra en dålig situation, oavsett just den aktörens roll i uppkomsten av den dåliga situationen (Miller 2001: 460; Young 2006: 126; jfr Nihlén Fahlquist 2009: 111). Aktören kan tillhöra en grupp av aktörer som, nu eller i förfluten tid, har bidragit till exempelvis klimatförändringar. Principen om gruppsolidaritet (jfr ovan) kan då rättfärdiga tillskrivandet av ansvar.

## Maskinen som expert

Den AI som finns idag är specialiserad på en enda uppgift. Ansiktsigenkänningsprogram, stöd för cancerdiagnoser inom sjukvården och rekryteringsprogram är exempel på olika tillämpningar som har det gemensamt att de är tränade på att känna igen en viss sorts mönster. Detta gör de med större precision ju fler ansikten, cancertumörer eller rekryteringar de har exponerats för, och man skulle kunna säga att de blir experter på sin enda uppgift. Kanske kan vi jämföra maskinen med den mänskliga experten när det gäller frågan om ansvar? En expert brukar definieras som någon med överlägsen specialistkunskap som kan tala med auktoritet inom ett väl avgränsat område, men inte på andra områden (Schudson 2006: 499; Turner 2003: 6), så i det avseendet finns en parallell. Skulle vi kunna överföra resonemang om mänskliga experters ansvar till ett resonemang om expert-maskinens ansvar?

När är experten ansvarig? Från ett idealtypiskt rationalistiskt perspektiv är experter i politiska beslutsprocesser neutrala leverantörer av kunskap, medan politiker gör värderingar och fattar beslut, som de sedan kan ställas till ansvar för. Detta synsätt, som också ligger till grund för den formella uppdelningen av roller i politiska beslutsprocesser, befriar experten från ansvar för de beslut som den levererade kunskapen bidrar till. Överfört till expert-maskinen skulle alltså inte heller den vara ansvarig för beslut som bygger på den kunskap som maskinen levererat. Det kan tyckas som att saken därmed borde vara avgjord, men åtminstone två aspekter komplicerar ansvarsfrågan.

Den ena komplikationen är att det rationalistiska synsättet är idealtypiskt (och formellt) och inte nödvändigtvis beskriver vad som faktiskt sker vid (politiskt) beslutsfattande. Ett konstruktivistiskt synsätt ifrågasätter förekomsten av neutral kunskap och pekar på vikten av att uppmärksamma den politiska roll som experten i praktiken kan spela. Dels kan experten konsulteras för att ge en rekommendation, och en rekommendation bygger alltid på någon form av

värdering. Dels kan experten i kraft av sin kognitiva auktoritet (Turner 2003: 23) avsiktligt eller oavsiktligt smyga in värderingar i det som kan förefalla vara neutral information. Forskares kognitiva, eller epistemiska, auktoritet i samhället ger dem stora möjligheter att påverka politiska beslut (Gundersen 2018: 54–55), och om experter de facto spelar en politisk roll, är det inte självklart att de ska vara befriade från ansvar. Om en expert-maskin på motsvarande sätt som en mänsklig expert ger rekommendationer, som exempelvis sker inom missbruksvården i Helsingborgs kommun (AI Lund 2020-01-29; MVTe 2020-01-21) men också i flera andra kommuner i Sverige (Svensson 2019: 27–32) och i Norden (Nordiska ministerrådet 2019: 18–20), eller till och med fattar beslut, som den så kallade handläggarroboten inom socialtjänsten i Trelleborg (Trelleborg 2015), skulle man alltså kunna hävda att även den borde ha ett ansvar.[2] Eller kan man det? Är det inte skillnad på maskiner och människor? Detta leder in på en andra aspekt som komplicerar ansvarsfrågan.

Den andra komplikationen är just att maskinen är en maskin. Även om maskinen bearbetar information och drar slutsatser som resulterar i en rekommendation, eller till och med ett beslut, gör den det på grundval av någon form av förprogrammerad logik. Det kan se ut som att maskinen fattar ett självständigt beslut, men den gör det på premisser som en människa (eller flera människor) har bestämt. Och till skillnad från den mänskliga experten, kan inte maskinen gå utöver sin uppgift att leverera svaret på den fråga som ställts, till exempel huruvida en viss person är berättigad till ett visst bidrag. Där den mänskliga handläggaren har möjlighet och förmåga att ta individuella hänsyn, genom att beakta faktorer som inte finns med i 'formeln', kan maskinen bara göra det den är programmerad att göra. Som vi såg i ansvarsdiskussionen ovan, kan möjligheten att göra annorlunda vara en förutsättning för ansvar. Att jämföra maskinen med en mänsklig expert är alltså inte helt adekvat för frågan om ansvar.

2 Huvudregeln är att automatiserade individuella beslut inte är tillåtna (GDPR art. 22.1). Med detta avses beslut med rättsliga följder (t.ex. diskriminering eller att en anställning sägs upp) eller som på liknande sätt i betydande grad påverkar en människa och som enbart grundas på automatiserad behandling. En människa måste fatta själva beslutet (på riktigt, inte bara symboliskt), men beslutsunderlaget kan tas fram av AI, algoritmer etc. Detta gäller både privata företag och offentlig sektor. Det finns dock ett antal undantag till denna huvudregel. Ett sådant undantag är om automatiserat beslutsfattande är tillåtet enligt nationell lag (GDPR art. 22.2). Den nya förvaltningslagen (2017:900) som började gälla 1 juli 2018 öppnar upp för automatiserat beslutsfattande (28 §). Rättsläget är dock inte helt klart, och olika aktörer gör olika tolkningar. Enligt SKR är automatiserat beslutsfattande tillåtet i staten, men inte inom kommuner och regioner, och SKR bedriver ett påverkansarbete för att automatiserat beslutsfattande ska vara lagligt även i kommuner och regioner (skr.se). I skrivande stund pågår en utredning (dir. 2020:10) som bland ser över hur detta kan införas i kommuner och regioner. Utredningens förslag väntas i december 2020.

## Om maskinen är autonom?

Maskiner kan alltså inte vara ansvariga, även om de i teknisk mening kan fatta beslut eller ge rekommendationer. Till skillnad från människor, saknar de självbestämmande och förmåga att reflektera över sina handlingar. Maskiner har heller ingen uppfattning om rätt och fel, och saknar därför förmåga att göra moraliska bedömningar. Utvecklingen av självlärande maskiner och autonoma system som agerar utan direkt mänsklig inblandning väcker dock frågor om huruvida maskiner kan och bör vara ansvariga. Autonomi, eller självbestämmande, brukar anföras som en förutsättning för ansvar. Det man inte själv har bestämt, kan man inte heller svara för, men om en maskin är autonom, kan den då också vara ansvarig för de beslut den fattar? Obemannade fordon, finansrobotar och robotar som kan ställa medicinsk diagnos är exempel på maskiner som ofta omnämns som autonoma system (EGE 2018), men kan vi säga att dessa maskiner är autonoma i den mening som förutsätts för ansvar?

Ordet autonomi kommer av grekiskans *autos* (själv) och *nomos* (lag) och kan uttydas som att den som är autonom, bestämmer sina egna lagar (Hakli & Mäkelä 2019: 264). Autonomi betyder alltså självstyre. En autonom stat eller region har friheten att styra sig själv, och en individ som är autonom har förmåga att agera och fatta beslut utan att kontrolleras av någon annan – men kan en maskin vara autonom? En fabriksrobot som dag efter dag, dygnet runt, utför sina uppgifter utan inblandning av någon människa, är den autonom? Och hur är det med självlärande system, som självkörande fordon eller spelet AlphaGo, som blir bättre och bättre på att utföra sina uppgifter ju mer de tränar på dem? Även om maskininlärning har effekten att fordonets eller spelets autonomi ökar, bygger denna autonomi på förprogrammerad kod (Burri 2016: 349), och kan vi då tala om autonomi i betydelsen självständig och oberoende? Det kan här vara på sin plats att påpeka att termen ’autonom’ används på olika sätt i olika sammanhang.

Den tekniska förståelse av autonomi som forskare inom AI och robotik använder kan uttryckas som maskinens förmåga att uppfatta omgivningen, att lära från sina erfarenheter och att agera oberoende av en extern operatör (Hakli & Mäkelä 2019: 264). Autonomi tolkas här relativt till mängden mänsklig inblandning medan roboten utför sin uppgift (Haselager 2005: 518). En robot är alltså inte autonom om en människa styr den medan den agerar. Däremot anses maskinen ha autonomi när den kan utföra en förprogrammerad uppgift utan inblandning från en människa (Burri 2016: 346; Jain & Pratihar 2010: 1). Även när en människa har bestämt vad maskinen ska göra, anses den i teknisk mening vara autonom så länge själva genomförandet av uppdraget sker utan mänsklig vägledning (Haselager 2005: 518).

Ur ett filosofiskt perspektiv däremot, är en robot som följer en rutin som bestämmer robotens handlingar inte autonom, utan följer bara de direktiv som en människa (eller flera människor) på förhand har satt upp (Burri 2016: 354).

En filosofisk, konceptuell förståelse av autonomi betonar kapaciteten att agera självständigt och att göra sina egna val (Haselager 2005: 519). Att vara autonom innebär att vara självbestämmande, oberoende och inte styrd av någon extern kraft. I Kants mening syftar autonomi på grundläggande villkor för den fria viljan, nämligen viljans kapacitet att följa moraliska lagar som den har givit sig själv (Gunkel 2017: 4). Det är således inte hur, utan varför, en uppgift utförs som är av intresse. För att vara autonom i filosofisk mening måste en maskin alltså ha förmåga att välja sina mål och att agera fritt utan mänsklig kontroll, varken under själva genomförandet av målet eller dessförinnan (Hakli & Mäkelä 2019: 265). Sådana förmågor hänger ihop med attityder och värderingar, principer och önskningar, och den självkontroll och det omdöme som är en förutsättning för att verkligen kunna göra det som krävs för att uppnå sitt mål (Hakli & Mäkelä 2019: 265), egenskaper som förknippas med agentskap (mer om det nedan).

Skillnaden mellan den ingenjörsmässiga och den filosofiska förståelsen av autonomi har delvis sin grund i olika temporala perspektiv. Det tekniska synsättet är inriktat på vad som är genomförbart just nu, medan det filosofiska synsättet är mer inriktat på framtiden (Burri 2016: 355) och på vad som är principiellt möjligt. Detta aktualiserar frågan om vilka maskiner som diskussionen om ansvar bör handla om: maskiner som finns just nu, eller alla maskiner, även sådana som kan komma att finnas i framtiden. Vi bevittnar nu en oerhört snabb utveckling av maskiner som lär sig själva utan att få nya instruktioner från dem som designar och programmerar dem. Några exempel är det ovan nämnda spelet AlphaGo, som är designat för att formulera sina egna instruktioner (Gunkel 2017); rekryteringsprogram som lär sig att hitta lämpliga kandidater genom att exponeras för tidigare rekryteringar (O'Neill 2016); och sociala medier som Facebook, som hela tiden anpassar vad som visas i varje användares flöde utifrån vars och ens interaktionsmönster (Bucher 2018). Självlärande maskiner av detta slag lär sig inte genom att få nya instruktioner från programmeraren, utan genom att interagera med omgivningen, med användarna och med olika situationer, och blir därmed allt mer autonoma i filosofisk mening (Laukyte 2017: 12).

När den omgivande miljöns påverkan ökar och programutvecklaren successivt förlorar kontroll, kan vi verkligen fråga oss vem som är ansvarig för vad maskinen gör, och inte minst vem som är ansvarig för konsekvenserna av maskinens beslut. När det gäller spelet AlphaGo kan konsekvenserna betraktas som relativt harmlösa (om man inte heter Lee Sedol och inte längre är en oslagbar Go-spelare), men konsekvenserna av diskriminerande rekrytering eller att var och en presenteras olika världsbilder i sociala medier är betydligt allvarligare. Vem, om någon, kan rimligen hållas ansvarig för dessa effekter? Bevittnar vi, som Matthias (2004) hävdar, en ansvarslucka – *a responsibility gap* – där det inte är möjligt att förutse maskinens beteende och därmed inte att hålla någon

ansvarig för vad maskinen gör? Eller kan en maskin, som utför sina uppgifter med en viss grad av autonomi vara ansvarig för vad den gör?

I en strikt kausal och momentan mening skulle man möjligen kunna säga att en autonom maskin är ansvarig (jfr Fischer 2007: 65; Hakli & Mäkelä 2019: 269). I det specifika ögonblick som den fattar ett beslut – till exempel att visa dig reklam för en viss produkt i ditt Facebook-flöde – orsakar maskinen, i det här fallet algoritmen, att du får se just det du får se. Men det är svårt att hävda att maskinen också skulle ha ett moraliskt ansvar för att du får se just denna produkt. Moraliskt (eller juridiskt) ansvar förutsätter en moralisk agent, och autonomi i teknisk mening eller en viss grad av autonomi i filosofisk mening är otillräckligt för moraliskt agentskap. Men om vi släpper tanken på de maskiner som existerar här och nu, och istället funderar mer principiellt över möjliga framtida maskiner som kan agera med ännu större autonomi: kan en maskin någonsin vara en moralisk agent?

## Om maskinen har moralisk agens?

Enligt ett instrumentalistiskt synsätt är det alltid människan som är ansvarig för vad ett teknologiskt system gör, hur oberoende, automatiskt eller interaktivt det än är, eftersom det alltid kommer att vara en produkt av mänskligt beteende, mänskliga institutioner och mänskliga beslut (Gunkel 2017: 3). Men självlärande system, med tillämpningar som exempelvis prognoser, datorspel eller mönsterigenkänning, och maskiner som har förmåga att förbättra sig själva – gör de inte något på egen hand och som är oberoende av mänsklig inblandning? Det finns de som hävdar att maskiner under vissa omständigheter kan ha moralisk agens, medan andra menar att det varken är möjligt eller önskvärt. Men vad innebär det att vara en moralisk agent?

Typiskt sett brukar en moralisk agent förstås som någon (eller möjligen något) som kan skilja på rätt och fel eller som kan hållas ansvarig för sina handlingar (Malle 2016: 244). För att vara en moralisk agent måste en aktör alltså vara medveten om en handlings moraliska natur (Held 1970: 97). Ofta anförda villkor på moralisk agens är olika kapaciteter såsom rationalitet, intentionalitet, förmåga till självreflexion, känslighet för moraliska skäl, förmåga att erfara känslor, social förmåga och att vara en person (*personhood*) (Hakli & Mäkelä 2019: 262). En moralisk agent måste också ha möjlighet att handla annorlunda (Fischer 2006: 65). Moraliskt agentskap förutsätter både handlingsfrihet och handlingsförmåga, annars kan individen inte på ett övertygande sätt prisas eller klandras för sitt beteende (Sondermann m.fl. 2018: 3; jfr Lundquist 1992: 41). Agentskap är emellertid i stor utsträckning villkorad av sociala praktiker och vanor, och inte ett resultat av fri vilja (Sondermann m.fl. 2018: 2).[3] Den

3 För en översikt över den fria viljans problem, se Fischer 1999.

dominerande uppfattningen är ändå att vuxna människor är moraliska agenter, medan icke-mänskliga agenter som djur, robotar och grupper inte är det (Hakli & Mäkelä 2019: 260, 265). Med den ökande betydelsen av artificiell intelligens och robotik, är det emerllertid viktigt att uppmärksamma också icke-mänsklig agens, inte bara därför att dessa intelligenta system utvecklas till att agera och att lära sig mer och mer självständigt (Sondermann m.fl. 2018: 4), utan också därför att maskiner som agenter utmanar hur vi vanligtvis ser på ansvar. Frågan är alltså om maskiner kan ha moralisk agens, och om de i så fall kan eller bör vara ansvariga för de beslut de fattar.

Vissa forskare menar att moralisk agens inte bör ses som något antingen – eller, utan som ett kontinuum på motsvarande sätt som vi inte har några problem med att tillskriva barn särskilda kompetenser i grader eller stadier (Sharkey 2017: 211). Sullins (2006) menar till och med att robotar kan vara fullständiga moraliska agenter.[4]

Utgångspunkten för Sullins resonemang är att medvetande inte är en förutsättning för moral, och givet denna utgångspunkt, skisserar han tre förutsättningar: 1) maskinen måste vara signifikant autonom, i teknisk bemärkelse, i förhållande till programmerare eller operatörer, 2) den måste kunna uppvisa ett beteende som kan uppfattas som en avsikt att göra gott eller ont, och 3) den måste bete sig på ett sådant sätt att beteendet kan uppfattas som att den har ansvar för någon annan moralisk agent (2006: 28–29). Även om det just nu inte finns några maskiner som kan uppvisa dessa beteenden, blir AI-utvecklare allt mer kapabla att genom programmering imitera mänskligt beteende. De förutsättningar som Sullins anför är rent funktionalistiska och innebär inte att maskinen är autonom, har intentioner eller är ansvarskännande i filosofisk mening, bara att den beter sig på ett sådant sätt att beteendet kan tolkas som att den agerar autonomt med intentioner och ansvarsfullt gentemot andra. Om exempelvis en vårdrobot beter sig på detta sätt, skulle beteendet kunna tolkas som att den har en övertygelse om att den ska bry sig om sina patienter, och om beteendet kan tolkas så, då kan vi enligt detta sätt att resonera tillskriva den moralisk agens (Sullins 2006: 29). Det skulle alltså vara tillräckligt att maskinen är moralisk i funktionell mening för att kunna tillskriva den ansvar, ungefär på samma sätt som företag ses om ”personer” i juridisk mening. Samma moraliska och legala överväganden skulle enligt detta resonemang vara möjliga att överföra till en AI eller robot som Google’s DeepMind eller IBM:s Watson (Gunkel 2017: 10).

Till detta kan läggas vår tendens att antropomorfiera saker, och att vi

4 En extrem position är att robotar, men inte människor, är moraliska agenter, en uppfattning som faktiskt en person har sagt sig ha. I en artikel från 2006 hävdar Joseph Emile Nadeau att den frihet som krävs för moraliskt agentskap förutsätter förmåga att tänka strikt logiskt och fullständigt tänka igenom alla skäl, och eftersom den mänskliga hjärnan saknar den förmågan, kan vi inte vara moraliska agenter (Sullins 2006: 27).

tenderar att reagera emotionellt på sociala robotar. En social robot är social på ett människolikt sätt och att interagera med den kan i många avseenden vara som att interagera med en annan människa. Flera empiriska studier har visat att mänskliga användare, som ett resultat av social interaktion, ger social ställning till datorer som liknar den som vi ger till andra människor. Till exempel har människor visat förmåga att empatisera med simulerad ”robotsmärta” (Gunkel 2017: 7–8). Men detta säger rimligtvis ingenting om *maskinens* moraliska agens. Det pekar snarare på hur sociala robotar suddar ut gränsen mellan moraliska subjekt och maskinen som ett rent instrument, vilket är kärnan i det problem som denna text tar sin utgångspunkt i.

Ett annat sätt att funktionalistiskt argumentera för att en maskin skulle kunna vara en agent är att jämföra med hur vi ser vi på gruppers agentskap. Liksom grupper, kan maskinen beskrivas som en icke-mänsklig agent, så om den uppfyller de villkor som vi ställer på gruppers agentskap, skulle även maskinen kunna erkännas som en agent, går resonemanget. Laukyte (2017) pekar på ett antal sådana krav, av vilka förmågan att agera rationellt och förmågan att interagera med sin miljö i sammanhanget är relativt oproblematiska. Dessa förmågor är utmärkande för de maskiner vi talar om här, nämligen robotar och algoritmer som bygger på artificiell intelligens. Mer utmanande är kravet på att vara en person (*personhood*). Vad det innebär att vara en person skiljer sig mellan olika kulturer och kontexter och är en kontroversiell fråga inom filosofi och juridik.[5] Relevant i detta sammanhang är hur en funktionalistisk syn på vad det innebär att vara en person kan göra det möjligt att argumentera för att en maskin kan erkännas som en person, om än en artificiell sådan, och därmed tillerkännas agentskap.

Laukyte utgår från ett performativt personbegrepp, utifrån vilket vi kan bedöma maskinen utifrån vad den gör eller kan göra. Vi har redan konstaterat att en maskin kan interagera med sin miljö, och detta kan nu kvalificeras till att handla såväl om den fysiska miljön som om den sociala miljön. Enligt det performativa personbegrepp som Laukyte använder, kan maskinen betraktas som en person om den kan agera i båda dessa miljöer: den fysiska världen och den sociala världens interaktioner. För att kunna agera i den sociala världen krävs kapacitet att operera inom ett system av krav och förväntningar och förmåga att interagera med människor och grupper inom etablerade konventioner. Om maskinen klarar detta, kan den enligt Laukyte (2017: 8–9) betraktas som en person, och uppfyller därmed även detta krav på agentskap. Detta performativt-relationella synsätt påminner om Turing-testet, som handlar om möjligheten att skilja på en dator och en människa.[6] Att vara en person

5 För en översikt, se Carrithers et al. 1985.

6 Det ursprungliga Turing-testet lanserades 1950 av matematikern och kryptoanalytikern Alan Turing, som satte upp ett imitationsspel där en utfrågare skulle avgöra om svaren kom från en man eller en kvinna. Turing ersatte sedan mannen (som det var) med en dator. Ansatsen att använda Turing-test för

handlar alltså om vad en agent kan göra i en social miljö och om den kan bete sig på ett sätt som överensstämmer med sociala förväntningar, och en funktionalistisk ansats gör det enligt Laukyte möjligt att låta komplexa sociotekniska system – som teknologier baserade på artificiell intelligens – ansvara för skador som de orsakar eller bidrar till (2017: 9).

Andra menar att ett teknologiskt system aldrig, hur oberoende, autonomt eller interaktivt det än är, kan vara en moralisk agent. Invändningar mot det funktionalistiska synsätt på moralisk agens som Sullins och Laukyte ger uttryck för betonar social förmåga och betydelsen av att vara en biologisk varelse och förmågan att erfara känslor. Maskinens brist på känslor skulle göra den till en moraliskt blind, artificiell psykopat (Gunkel 2017: 10), medan biologin sägs vara grunden för anknytning och empati, för utveckling av komplexa sociala relationer, för förmåga att förstå och förutse andras handlingar, och för att erfara social smärta. Detta sägs leda till att människor – och i stor utsträckning även andra däggdjur – har en inneboende känsla för rättvisa (Sharkey 2017: 211). Att bry sig om sig själv och andra är grunden för utvecklingen av moral, men robotar bryr sig inte om sitt självbevarande eller att undvika smärta, och ännu mindre andras smärta (Sharkey 2017: 211). Enligt det biologiska argumentet beror det på att maskinen inte har en kropp på samma sätt som levande varelser. Man skulle kunna ta bort delar av roboten utan att den känner någon smärta eller oro, och ännu mindre skulle den bry sig om skada eller smärta som en människa drabbas av. Och även om roboten kan programmeras att bete sig som om den brydde sig om att skydda sig själva eller andra, skulle det bara vara möjligt genom mänsklig intervention (Sharkey 2017: 212).

Den mänskliga inblandningen återkommer i olika argument mot att tillskriva agentskap till maskiner. En sådan invändning är att det är logiskt inkorrekt att tillskriva agentskap, och därmed ansvar, till något som enbart är ett objekt under mänsklig kontroll (Gunkel 2017: 3). Det skulle också vara moraliskt problematiskt, då det skulle göra det möjligt att skylla våra misstag på datorer (Gunkel 2017: 3).

Som vi har sett, tas ofta den tekniska utvecklingen som intäkt för att maskiner bör kunna betraktas som agenter. När maskiner i allt större utsträckning klarar av sådant som brukar förknippas med mänskligt agentskap, dras slutsatsen att också maskiner kan tillskrivas agentskap med allt vad det innebär av rättigheter och ansvar. En invändning mot att på detta sätt utgå från kapaciteter tar fasta på hur kapaciteterna har uppstått. Även om robotar skulle ha alla kapaciteter som krävs för moraliskt agentskap, skulle enligt Hakli & Mäkelä (2019) deras historia beröva dem autonomi på ett sätt som underminerar ansvar. Även om maskiner kan ha kapaciteter som intentionalitet och rationalitet, är det tveksamt om de skulle kunna ha förmåga till självreflexion, känslor,

att tillskriva *personhood* till robotar istället för att helt enkelt säga att de gör något, har kritiserats (se till exempel Bringsjord 1992). Turing beskrivs ofta som den artificiella intelligensens fader.

och normativ förståelse. Och även om alla dessa kapaciteter kommer att kunna att implementeras i robotar, skulle detta inte göra robotarna till moraliska agenter, eftersom det spelar roll hur kapaciteterna har uppnåtts (Hakli & Mäkelä 2019: 264). Liksom Laukyte ovan, liknar Hakli & Mäkelä roboten vid grupper, men de kommer till en annan slutsats. De pekar på att kollektiva agenter, som företag och andra grupper, inte har den kontroll över sina handlingar som krävs för moraliskt ansvar, utan kontrollen ligger i de individuella medlemmarnas händer (2019: 266–267). På samma sätt är robotens agentskap beroende av sin skapare. Utan en historia av att själv erövra kapaciteter, värderingar och attityder, är robotens agentskap inte autentiskt, och det går därför inte att tillskriva roboten agentskap. Manipulationen undergräver autonomin, och därmed möjligheterna att hålla roboten ansvarig (2019: 268). Medan Laukyte (2017: 3) hävdar att vi inte behöver bry oss om maskinens ”inner workings”, det vill säga hur den har kommit fram till vad den ska göra, utan bara vad den faktiskt gör, menar Hakli & Mäkelä att det är just ”the inner workings” som är relevant för om vi kan tillskriva en maskin moralisk agens.

## Maskinen som husdjur

I diskussioner om huruvida framtida robotar ska ha rättigheter är det vanligt att jämföra maskinen med ett djur (Sullins 2006; Gunkel 2017; Köhler m.fl. 2018). Jämförelsen mellan maskin och djur vilar på antagandet att det inte är någon signifikant moralisk skillnad mellan en maskin och ett djur (Hogan 2017: 29). En ofta anförd analogi är den mellan sociala robotar och familjehunden. Familjehunden agerar utifrån sina egna önskningar, men är också tränad att utföra vissa uppgifter i vissa situationer. På motsvarande sätt försätts den sociala roboten eller drönaren i särskilda omständigheter i vilka den sedan agerar autonomt (Köhler m.fl. 2018: 57–58, 63). Programmeringen av roboten är här att jämföra med dresserandet av hunden (Johnson & Verdicchio 2018: 296).

Den huvudsakliga invändningen mot husdjursanologin är att den är missledande. Djur har medvetande, vilket inte roboten har, och djur har förmåga att uppleva känslor och kan lida, vilket roboten inte heller kan (Johnson & Verdicchio 2018: 294). En invändning mot denna invändning är att framtida robotar skulle kunna ha medvetande och lida. Djur får moralisk status från sin kapacitet att lida, och om robotar har kapacitet att lida, bör de också ha moralisk status (Johnson & Verdicchio 2018: 295). Men så länge maskiner inte har denna förmåga, leder jämförelsen fel. Djurens fysiska existens är ett skäl till deras moraliska status (jfr diskussionen om biologi ovan), medan maskinens eventuella moraliska status vilar på deras mentala kvaliteter: autonomi, intelligens och rationalitet (Hogan 2017: 31).

Andra argument för att robotar och djur inte är jämförbara utgår från hur deras relation till människor skiljer sig åt. För maskiner är frågan om huruvida

de ska behandlas etiskt, men när det handlar om djur, är frågan på vilket sätt de ska behandlas etiskt (Hogan 2017: 29). Vidare, i diskussioner om maskinens moraliska status handlar frågan om huruvida de ska betraktas som moraliska agenter, det vill säga om de ska kunna hållas ansvariga för skador som de orsakar, medan det för djur handlar om huruvida de ska betraktas som moraliska patienter (eng. *patient*), det vill säga, huruvida de förtjänar moralisk hänsyn (Hogan 2017: 30). En standarduppfattning inom etiken är att alla agenter också är patienter, medan alla patienter inte nödvändigtvis är agenter. Familjehunden förtjänar etisk hänsyn trots att den inte tar ansvar för sina handlingar. Robotetiken besvarar frågan om robotens patientskap med hänvisning till robotens agentskap, medan djuretiken erkänner moraliskt patientskap till varelser som vanligtvis *förnekas* moraliskt agentskap (Hogan 2017: 30–31). Djur och maskiner är således inte jämförbara när det gäller frågan om moralisk agens.[7]

## Om maskinen är medveten?

Men om maskinen är medveten, då borde den väl kunna vara ansvarig för beslut som den fattar?[8] Ja, givet att den också har adekvat kunskap om orsaksförhållanden och moral, och att den har möjlighet att agera annorlunda – det vill säga de krav på ansvar som vi ställer på mänskliga beslutsfattare – då skulle medvetande så att säga putta den över gränsen och göra maskinen till ett ansvarstagande subjekt. Medvetna maskiner är, om något, förknippat med science fiction, men det finns ändå tänkare, exempelvis inom transhumanismen, som tror att den snabba tekniska utvecklingen inom AI kan leda till att maskiner så småningom blir medvetna.

En avgörande skillnad verkar finnas mellan å ena sidan biologiska naturalister, som hävdar att förmågan att vara medveten är unik för biologiska organismer, och å andra sidan teknologiska optimister, som menar att medvetande är *substrate independent*, alltså oberoende av vilket material någonting är

7 Detta förhindrar inte att det ändå kan vara relevant att jämföra robotar med djur. Det kantianska argumentet att hur vi behandlar djur har betydelse för vad det gör med oss som människor kan utvidgas till att också gälla hur vi behandlar robotar: Om vi behandlar robotar (och djur) illa, blir vi inhumana människor. Vår behandling av robotar (och djur) kan förstärka mänskligt beteende som är moraliskt korrekt eller inkorrekt. Att behandla robotar (och djur) på ett grymt sätt gör oss känslomässigt avtrubbade och kan underminera den empati vi har för andra människor. Men vi vet inte om det finns en sådan överföringseffekt när det gäller djur, och om den finns, hur kraftfull den i så fall är. Och överfört till robotar blir argumentet ännu svagare, eftersom robotar inte lider. Eftersom roboten inte lider när någon slår den, är det inte adekvat att säga att detta beteende är grymt. Inte heller denna analogi mellan maskinen och djuret håller, eftersom maskinen inte lider (Johnson & Verdicchio 2018: 298). Men att ge robotar rättigheter kan ändå ha betydelse för vilka vi blir som människor, då roboten kan tjäna som en etisk förebild för människan (Levy i Hogan 2017: 31).

8 Det är inte helt lätt att definiera medvetande, men känslor, inre mentalt liv, inre erfarenheter och subjektiv erfarenhet ringar in vad det handlar om (jfr Schneider 2019: 16; Tegmark 2017: 315) och fångar på ett ungefär in vad medvetande är. För en mer ingående diskussion, se t.ex. Blackmore & Troscianko 2018.

uppbyggt av (jfr Schneider 2019: 17–26; Tegmark 2017: 315). Om och när människor utvecklar sofistikerade generella AI:ar som flexibelt skiftar mellan olika intellektuella uppgifter och som överträffar människans förmåga att resonera, så kommer dessa att ha medvetande, enligt denna position. Då – men inte förr – skulle vi kunna tillskriva maskiner ansvar.

Vi kan förstås inte veta om medvetna maskiner någonsin kommer att utvecklas och hur långt in i framtiden en sådan utveckling i så fall ligger (Bostrom 2014: 18–21). Än mindre kan vi veta hur vi kan veta att en maskin är medveten. Om vi accepterar slutsatsen att medvetande är en förutsättning för att vara ett ansvarstagande subjekt, är det en avgörande fråga om vi kan veta att maskinen verkligen är medveten. Även om maskinen beter sig som om den är medveten, behöver det inte innebära att den också är det. Den – eller de som utvecklar och säljer dem – kanske bara vill få oss att tro det. Medvetande är inte en oundviklig konsekvens av teknisk utveckling, och även om framtida maskiner skulle kunna vara medvetna, skulle självförbättrande system kanske utveckla bort medvetandet – kanske för att slippa ta ansvar? Svårigheten att verkligen avgöra om en maskin är medveten föranleder filosofen Susan Schneider att förespråka en försiktighetsprincip: från ett etiskt perspektiv är det bäst att anta en sofistikerad AI kan vara medveten (2019: 149).

## Är maskinen ett hinder för ansvar?

Om nu slutsatsen blir att maskiner – såvida de inte är medvetna – *inte* kan vara ansvariga, bör vi fundera över vilka uppgifter som vi ska överlämna till maskiner, hur effektivt och kostnadsbesparande det än kan vara. Det kan till exempel, som Sharkey (2017) hävdar, vara rimligt att begränsa maskiners roll i situationer där maskinen kan fatta beslut som har konsekvenser för människors liv och välbefinnande: stridsrobotar eller robotpoliser kan orsaka död eller allvarlig skada, robotlärare kan behöva fatta beslut om att straffa eller begränsa barn, robotar inom äldrevården kan vara tvungna att fatta beslut om att dela information om de äldre eller att hindra dem från att göra något farligt. Sådana beslut innebär moraliska bedömningar av sociala situationer och bör enligt Sharkey (2017: 214) göras av människor, som både har förmåga att fatta ansvarsfulla beslut (framåtblickande ansvar) och kan ställas till svars för dåliga beslut (bakåtblickande ansvar).

Även om försök görs att, utifrån vad allt komplexare robotar och algoritmer kan göra, argumentera för att maskiner kan tillskrivas ansvar, kan vi konstatera att det inte är helt enkelt att göra det på ett övertygande sätt. Även om maskinerna beter sig som mänskliga aktörer och kan få oss att tro att de verkligen är autonoma agenter, är det svårt att komma ifrån att de är skapade av människor och att allt som de gör, gör de därför att en människa, eller flera människor tillsammans, har bestämt att de ska göra just det de gör. Medvetande verkar vara

förutsättningen för att kunna ta ansvar på ett meningsfullt sätt. Ändå är frågan om maskiners ansvar högst aktuell. Den snabba tekniska utvecklingen av artificiell intelligens och självlärande maskiner väcker onekligen många frågor om vem som är ansvarig för beslut som vi konfronteras med dagligen. Vi ska nu, avslutningsvis, närma oss frågan om maskinen kan utgöra ett hinder för ansvar genom att återkomma till debatten om en ansvarslucka, som nämndes i förbigående tidigare.

Diskussionen om en ansvarslucka tar avstamp i utvecklingen av autonoma maskiner. Det handlar här alltså om autonomi i den tekniska betydelsen att utföra en förprogrammerad uppgift utan inblandning från en människa. Självlärande system gör också vad de är programmerade att göra, men exakt hur de gör det och exakt vart det leder kan inte designern veta. För att återvända till spelet AlphaGo, så är det utformat för att formulera sina egna instruktioner. Ingenjörerna som utformade och byggde det har liten eller ingen idé om vad det kommer att göra när det körs. Självlärande system är avsiktligt utformade för att göra saker som deras programmerare inte kan förutse eller helt kontrollera (Gunkel 2017: 6). Ansvarsluckan uppkommer alltså därför att ingen har tillräcklig kontroll över vad maskinen gör. Matthias (2004: 177) menar att autonoma, självlärande maskiner skapar en situation där tillverkaren eller operatören *i princip* inte är kapabel att förutse den framtida maskinens beteende och därför inte kan hållas moraliskt ansvarig för den.

Ett exempel som illustrerar svårigheten med ansvar är den mobila leksaksroboten Aibo. Den har förmåga att lära sig, till exempel en ny gångstil, när den inbyggda algoritmen för att gå visar sig inte vara optimal för tjocka persiska mattor. Efter litet experimenterande upptäcker den att batteritiden kan förlängas om den galopperar, vilket minskar friktionen till underlaget. Antag att roboten när den galopperar omkring råkar kollidera med ett litet barn, som ramlar omkull och skadas. Vem är ansvarig? Tillverkaren? Varför, i så fall? Barnets föräldrar för att ha placerat leksaksroboten i lägenheten? Eller är detta en oförutsebar utveckling, som skedde på grund av robotens adaptiva förmågor, så att ingen kan sägas vara ansvarig (Matthias 2004: 177)? Om detta är en ansvarslucka, är det något som behöver hanteras i lagstiftning och moralisk praktik (Matthias 2004: 183), men det finns invändningar mot resonemanget.

Köhler och kolleger (2018) menar att det inte finns någon ansvarslucka, och lyfter fram två skäl. För det första pekar de på att det visserligen inte är någon som kontrollerar autonoma maskiner medan de opererar. Däremot är kontroll inblandad när maskinen försätts i olika omständigheter i vilka man vet att den kommer att agera autonomt (2018: 57–59). Jämförelsen med dresserade djur kommer här tillbaka, men här för att illustrera att djuret *inte* kan vara ansvarigt. För det andra, om det bara handlar om att det är svårare att *veta* vem som är ansvarig eller i vilken grad, finns inget problem (2018: 55; jfr Young 2006). Att det är svårt att reda ut vem som är ansvarig för vad skapar inte en lucka.

Det implicerar bara att många personer behöver hållas ansvariga, vilket snarare handlar om oklart ansvar (Köhler m.fl. 2018: 63–65).

Men vad skulle det innebära om maskiner *kan* vara ansvariga? Även om nu existerande maskiner inte har sådana egenskaper som krävs för ansvar, kan vi inte veta om framtida maskiner skulle kunna ha det. Vad skulle en sådan maskin göra med ansvaret? Såvida inte kännande maskiner utvecklas, förefaller bakåtblickande och straff föga meningsfullt. Å andra sidan har vi precis konstaterat att medvetande är en förutsättning för ansvar, och då torde det vara rimligt att maskinen också uppfattar straffet som ett straff. Men *om* medvetna maskiner skulle bli verklighet, är det inte helt otänkbart att det är maskinerna, och inte vi människor, som avgör ansvarsfrågor. När det gäller de icke-kännande maskiner vi har omkring oss idag, och sannolikt under lång tid framöver, är det inte särskilt fruktbart att fundera över hur vi lämpligen straffar maskinen för klandervärda handlingar den fattar beslutar om. Det är människor som måste bära ansvaret för vad maskinerna gör. *Vilka* människor som ska bära detta ansvar och hur ansvaret ska fördelas mellan dem är en fråga som får behandlas i ett annat sammanhang.

Maskiner som fattar beslut bidrar definitivt till att försvåra fördelningen av ansvar, men för att kunna hävda att maskinen är ansvarig, krävs att den är medveten, med allt vad det innebär av att kunna känna smärta och uppleva andra känslor. Och dit är det sannolikt en bit på väg.

## Referenser

Adam, Barbara & Groves, Chris, 2011. ”Futures tended: Care and future-oriented responsibility”, *Bulletin of Science, Technology & Society* 31(1), s. 17–27.

AI Lund, 2020-01-29. Presentation på workshop. http://ai.lu.se/events/?event=ai-lund-fika-till-fika-workshop-om-ai-i-offentlig-sektor, http://ai.lu.se/tv/ai-i-offentlig-sektor/.

Albertsen, Andreas, 2015. ”Feiring’s concept of forward-looking responsibility: a dead end for responsibility in health care”, *Journal of Medical Ethics* 41(2), s. 161–164.

Blackmore, Susan & Troscianko, Emily T., 2018. *Consciousness.* 3:e upplagan. Abingdon, Oxon and New York, NY: Routledge.

Bostrom, Nick, 2014. *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press.

Bringsjord, S., 1992. *What robots can and can’t be.* Dordrecht: Kluwer Academic.

Bucher, Taina, 2018. *If…Then: Algorithmic Power and Politics.* Oxford: Oxford University Press.

Burri, Thomas, 2016. ”The Politics of Robot Autonomy”, *European Journal of Risk Regulation* 7(2), s. 341–360.

Carrithers, Michael (red.), 1986. *The category of the person: Anthropology, philosophy, history.* Cambridge: Cambridge University Press.

EC Communication, 2018. ”Artificial Intelligence for Europe”, Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, COM (2018) 237.

EGE, 2018. *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems.* European Commission: European Group on Ethics in Science and new Technologies.

EP, 2017. European Parliament resolution ”Civil Law Rules on Robotics”, P8_TA(2017)0051.

EU-kommissionen, 2019. *Ethics Guidelines for Trustworthy AI.* High-Level Expert Group on Artificial Intelligence.

Feinberg, Joel, 1968. “Collective responsibility”, *The Journal of Philosophy* 65(21), s. 674–688.

Feinberg, Joel, 1970. ”Collective responsibility”, i May, Larry & Hoffman, Stacey (red.), *Collective Responsibility: Five Decades of Debate in Theoretical and Applied Ethics.* Lanham, MD: Rowman & Littlefield Publishers Inc.

Fischer, John Martin, 1999. ”Recent work on moral responsibility”, *Ethics* 110(1), s. 93–139.

Fischer, John Martin, 2007. ”Free will and moral responsibility”, i Copp, David (red.), *The Oxford Handbook of Ethical Theory*, s. 321–354.

Fischer, John Martin & Ravizza, Mark, 1998. *Responsibility and control: A theory of moral responsibility.* Cambridge: Cambridge University Press.

Gilbert, Margaret, 2006. ”Who is to blame? Collective moral responsibility and its implications for group members”, *Midwest Studies in Philosophy* 30(1), s. 94–114.

Gundersen, Torbjørn, 2018. ”Scienstists as experts: A distinct role?”, *Studies in History and Philosophy of Science* 69, s. 52–59.

Gunkel, David J., 2017. “Mind the gap: responsible robots and the problem of responsibility”, *Ethics and Information Technology,* https://doi.org/10.1007/s 10676-017-9428-2.

Hakli, Raul & Mäkelä, Pekka, 2019, “Moral responsibility of robots and hybrid agents”, *The Monist* 102, s. 259–275.

Haselager, Willem F. G.,2005, ”Robotics, philosophy and the problems of autonomy”, *Pragmatics & Cognition* 13(3), s. 515–532.

Hart, David Robert, 2018. “Saudi Arabia's robot citizen is eroding human rights”, *Quartz,* 2018-02-18: https://qz.com/1205017/ saudi-arabias-robot-citizen-is-eroding-human-rights/.

Held, Virginia, 1970. ”Can a random collection of individuals be morally responsible?”, i May, Larry & Hoffman, Stacey (red.), *Collective responsibility: Five decades of debate in theoretical and applied ethics.* Lanham, MD: Rowman & Littlefield Publishers Inc.

Hogan, Katharyn, 2017. ”Is the machine question the same question as the animal question?”, *Ethics of Information Technology* 19, s. 29–38.

HRW 2015: Human Rights Watch, 2015. *Mind the Gap: The Lack of Accountability for Killer Robots.* Harward Law School: International Human Rights Clinic.

Jain, Lakhmi C. & Pratihar, Dilip Kumar, 2010. *Intelligent Autonomous Systems.* Berlin/Heidelberg: Springer.

Johnson, Deborah G. & Verdicchio, Mario, 2018. ”Why robots should not be treated like animals”, *Ethics and Information Technology* 20(4), s. 291–301.

Kane, Robert, 2002. ”Responsibility, reactive attitudes, and free will: Reflections on Wallace’s theory”, *Philosophy and Phenomenological Research* 64(3), s. 693–698.

Köhler, Sebastian, Roughley, Neil & Sauer, Hanno, 2017. “Technologically blurred responsibility? Technology, responsibility gaps and the robustness of our everyday conceptual scheme”, i Ulbert, Cornelia, Finkenbusch, Peter, Sondermann, Elena & Debiel, Tobias (red.), *Moral Agency and the Politics of Responsibility*. London: Imprint Routledge.

Lanteri, Alessandro, 2009. ”Judgements of intentionality and moral worth: Experimental challenges to hindriks”, *The Philosophical Quarterly* 59(237), s. 713–720.

Laukyte, Migle, 2017. ”Artificial agents among us: Should we recognize them as agents proper?”, *Ethics of Information Technology* 19, s. 1–17.

Lundquist, Lennart, 1992. *Förvaltning, stat och samhälle*. Lund: Studentlitteratur.

Malle, Bertram F., 2016. ”Integrating robot ethics and machine morality: the study and design of moral competence in robots”, *Ethics of Information Technology* 18, s. 243–256.

Matthias, Andreas, 2004. “The responsibility gap: Ascribing responsibility for the actions of learning automata”, *Ethics and Information Technology* 6, s. 175–183.

May, Larry & Hoffman, Stacey (red.), 1991. *Collective responsibility: Five decades of debate in theoretical and applied ethics*. Lanham, MD: Rowman & Littlefield Publishers Inc.

McGary, Howard, 1986. ”Morality and collective liability”, i May, Larry & Hoffman, Stacey (red.), *Five Decades of Debate in Theoretical and Applied Ethics*. Lanham, MD: Rowman & Littlefield Publishers Inc.

Miller, David, 2001. ”Distributing responsibilities”, *The Journal of Political Philosophy* 9(4), s. 453–471.

MVTe: Mötesplats välfärdsteknologi och e-hälsa, konferenspresentation, 2020-01-21. https://www.mvte.se/sv/activity/kan-missbruksvarden-forbattras-med-hjalp-av-ai/.

Nihlén Fahlquist, Jessica, 2009. ”Moral responsibility for environmental problems–Individual or institutional?”, *Journal of Agricultural and Environmental Ethics* 22(2), s. 109–124.

Nordiska ministerrådet, 2019. *Nordiske kommuners arbeid med kunstig intelligens*. Ulf Andreasson & Truls Stende.

O’Neill, Cathy, 2016. *Weapons of Math Destruction*. New York: Crown.

Quora, https://sv.quora.com/Kan-en-AI-ber%C3%A4kna-och-f%C3%B6ruts%C3%A4ga-progressiv-utveckling.

Rovane, Carol, 2014. ”Forward-Looking Collective Responsibility: A Metaphysical Reframing of the Issue”, *Midwest Studies In Philosophy* 38(1), s. 12–25.

Russell, Stuart, 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking Press.

Schneider, Susan, 2019. *Artificial You: AI and the Future of Your Mind*. Princeton: Princeton University Press.

Schudson, Michael, 2006. ”The trouble with experts – and why democracies need them”, *Theor Soc* 35(5/6), s. 491–506.

Sharkey, Amanda, 2017. ”Can robots be responsible moral agents? And why should we care?”, *Connection Science* 29(3), s. 210–216.

Sellers, Mortimer, 2007. ”An Introduction to the Value of Autonomy in Law”, i Sellers, Mortimer (red.), *Autonomy in the Law*. Springer.

SKR, 2018. ”Automatiserat beslutsfattande i den kommunala förvaltningen”. Stockholm: Sveriges kommuner och Landsting.

Sneddon, Andrew, 2005. “Moral responsibility: The difference of strawson, and the difference it should make”, *Ethical Theory and Moral Practice* 8(3), s. 239–264

Sondermann, Elena, Ulbert, Cornelia & Finkenbusch, Peter, 2017. “Introduction: Moral agency and the politics of responsibility”, i Ulbert, Cornelia, Finkenbusch, Peter, Sondermann, Elena & Debiel, Tobias (red.), *Moral agency and the politics of responsibility.* London: Imprint Routledge.

Sullins, John P., 2006. “When is a robot a moral agent?”, *International Review of Information Ethics* 6(12), s. 23–36.

Svensson, Lupita, 2019. ”’Tekniken är den enkla biten’: Om att implementera digital automatisering i handläggningen av försörjningsstöd”, *Research report in social work 2019:12*. Lund: Socialhögskolan, Lunds universitet.

Talbert, Matthew, 2008. ”Blame and responsiveness to moral reasons: Are psychopaths blameworthy?”, *Pacific Philosophical Quarterly* 89(4), s. 516–535.

Techworld, https://techworld.idg.se/2.2524/1.722934/vd-rost-fejk-ai.

Tegmark, Max, 2017. *Life 3.0: Being human in the age of Artificial Intelligence.* Penguin Random House.

Thompson, Dennis F., 1987. *Political ethics and public office*. Cambridge, MA, och London: Harvard University Press.

Trelleborg, 2015. https://www.trelleborg.se/sv/omsorg-hjalp/hjalp-i-hemmet/for-en-smartare-valfard/

Turing, Alan M., 1950. ”Computing machinery and intelligence”, *Mind* 59, s. 433–460.

Turner, Stephen P., 2003. *Liberal Democracy 3.0*. London, Thousand Oaks & New Delhi: Sage Publications.

Voister, https://www.voister.se/artikel/2019/04/en-ai-vi-forstar/.

Wahlberg, Lena, 2019. ”Skäligt straffansvar för misstag inom medicinsk vård och behandling”, *Nordisk Socialrättslig Tidskrift*, särtryck ur NST, s. 19–20.

Young, Iris Marion, 2006. ”Responsibility and global justice: A social connection model”, *Social Philosophy and Policy* 23(1), s. 102–130.