Big Data og samfunnsforskning: Nye muligheter og etiske utfordringer

ENGLISH TITLE: *Big Data and social research: New possibilities and ethical challenges*

BERNARD ENJOLRAS
bernard.enjolras@samfunnsforskning.no

I KJØLVANNET AV digitaliseringens fremvekst i alle deler av samfunnet, har en teknologisk revolusjon som vil prege både samfunnet og samfunnsforskningen funnet sted. Mengden av tilgjengelige digitale data har eksplodert de siste årene. Det dreier seg om hverdagslige statusoppdateringer på Facebook, videoer lagt ut på YouTube og Twittermeldinger som er tilgjengelige for alle som vil lese dem. Det handler også om data fra kjøpstransaksjoner, søkemotorer og andre digitaliserte transaksjoner i offentlig sektor, helsevesen, skoleverk osv. Vi snakker om *Big Data* – et moteord som antageligvis vil bli erstattet av en ny betegnelse i nærmeste fremtid, men også betegnelsen på en utvikling som har kommet for å bli.

Begrepet Big Data er en samlebetegnelse for data som er av et slikt omfang at de krever mer enn vanlig datakraft for å samles inn, lagres og analyseres. Begrepet brukes ofte ikke bare for å betegne selve dataene, men også for å beskrive de nye problemstillingene slike data reiser, både teknisk, juridisk og etisk. Felles for Big Data er at de innebærer en registrering av faktiske handlinger, interaksjoner og transaksjoner koblet til individer.

I denne forskningskommentaren skal jeg drøfte nærmere hvilke etiske implikasjoner Big Data har for samfunnsforskning, men også hvilke muligheter som ligger i dette fenomenet for samfunnsforskningens utvikling. Jeg vil først argumentere for at samfunnsforskningens fremtid er avhengig av å benytte nye databehandlingsteknologier og nye datatyper forbundet med «*the Big Data turn*». For det andre vil jeg drøfte hvordan Big Data utfordrer vårt private liv og vårt personvern, også når det gjelder bruk av Big Data for samfunnsforskningsformål. Til slutt vil jeg diskutere dagens personvernsregulering og dens konsekvenser for bruk av Big Data til samfunnsforskningsformål. Jeg argumenterer for å overlate et større etisk ansvar til forskerne og forskningsmiljøene i stedet for dagens tvetydige, ineffektive og hemmende lovgivning.

TIDSSKRIFT FOR SAMFUNNSFORSKNING
VOL 55, NR 1, 80–89

BIG DATA OG SAMFUNNSFORSKNINGS FREMTID

Mengden av tilgjengelige data har eksplodert i løpet av de siste årene. Grunnen til denne utviklingen er en teknologisk revolusjon kjennetegnet av en rekke innovasjoner knyttet til overgangen til Web 2.0 (O'Reilly 2005), det vil si fra statiske websider til interaktive web- tjenester. Disse innovasjonene, parallelt med utviklingen av mobile internetteknologier, har dannet grunnlaget for flere innovasjoner, blant annet knyttet til tingenes Internett hvor ulike maskiner og apparater inkorporerer digitale elementer som er koblet mot Internett. Maskin-til-maskin-teknologi (M2M) innebærer å benytte fastnett, mobilnett eller trådløse nettverk for å kople sammen enheter og terminaler. Disse mulighetene anvendes i en rekke ulike bransjer som transport, kapitalforvaltning, vann- og strømleveranser og helsevesen. I tillegg til data generert av web-baserte tjenester, søkemotorer og sosiale medieplattformer, genererer internettkoblede enheter (mobile telefoner, GPS, bompasseringer osv.) data fra kunder og operasjoner som kontinuerlig blir lagret og analysert. Det er mulig å skille mellom fem ulike typer Big Data:

- *Web og sosiale medier data* som består av klikkstrøm og oppdateringer fra Facebook, Twitter, LinkedIn og blogger.
- *Maskin-til-maskin data (M2M)* som refererer til teknologier som muliggjør at elektroniske systemer (trådløse eller ikke) kommuniserer med hverandre. M2M-kommunikasjoner danner det såkalte «tingenes Internett», hvor maskiner utstyrt med ulike sensorer produserer meningsfull informasjon (som for eksempel GPS-posisjoner) som kan lagres for videre analyse.
- *Big transaksjonsdata* som består av helsejournaler, telekommunikasjonslogger, kundefakturering osv. Disse dataene inneholder også viktige metadata som kan utfordre personvernet. Metadata er informasjon (data) som beskriver dataene. Metadata (for eksempel et brukernavn, e-post eller IP-adresse) er avgjørende for å kunne koble ulike typer strukturerte elle ustrukturerte data sammen og dermed identifisere enkeltpersoner, samt samle en mengde informasjon om denne personen.
- *Biometriske data* er knyttet til automatisk identifisering av personer basert på anatomiske kjennetegn. Anatomiske data er generert gjennom lagring av individuelle fysiske kjennetegn som fingeravtrykk, iris, retina, ansikt, stemmemønster, DNA.
- *Menneskegenererte data* består av data som for eksempel samtaletapping, e-poster, surveydata, elektroniske helseregistre osv. som er produsert i ulike sammenhenger og organisasjoner.

Overgangen fra Web 1.0 til Web 2.0 har blitt drevet av og har drevet utviklingen av Big Data-teknologier. For å håndtere den stadig økende mengden av data har selskaper som Yahoo, Google, Amazon og Facebook utviklet nye modeller for datalagring basert på distribuert databehandling (*distributed computing*). Isteden for å øke beregnings- og lagringskapasitet ved å utvikle stadig større datamaskiner, muliggjør distribuert databehandling økning i beregnings- og lagringskapasitet ved å legge til flere beregningsenheter koblet sammen i et nettverk

(*cluster*). Distribuert databehandling innebærer at flere tusen datamaskiner koblet sammen i et nettverk jobber sammen om de samme oppgavene. Den best kjente konkretisering av distribuert databehandling er de store internettselskapenes datafarmer som består av flere tusener dataservere.[1]

Big Data har mange eksisterende og potensielle bruksområder innen onlinetjenester, digital markedsføring, svindelavsløring, risikostyring, helse, offentlig sektor osv. Big Data kombinert med maskinlæringsalgoritmer er basis for en rekke anvendelser i alle samfunnssektorer som predikerer individenes atferd. Et kjennetegn ved maskinlæringsalgoritmer er at deres ytelse øker med datamengden tilgjengelig for både trening og prediksjon. Teknologioptimister anser Big Data som løsningen som gjennom treffsikre prediksjoner vil bidra til effektivisering av ulike funksjoner i offentlig sektor, helse og næringsliv, og til bekjempelse av kriminalitet og terror. *Predictive Analytics* (Siegel 2013) er et område av datamining som har til hensikt å trekke ut informasjon fra data og bruke den til å forutse trender og atferdsmønstre. Datamining består i å analysere en stor mengde data ved hjelp av ulike statistiske og maskinlæringsmetoder for å finne mønstre. Resultater fra dataminingsprosesser (de estimerte modellene) kan anvendes til prediksjon, for eksempel å identifisere mistenkte for kriminelle handlinger.

Fra et slikt perspektiv kan også Big Data anses som et nytt verktøy for samfunnsforskningen. En rekke forfattere har uttrykt bekymring for at samfunnsvitenskapene risikerer å bli utryddet på sikt hvis vi overlater Big Data til andre.

Ifølge Mayer-Schönberger og Cukier (2013) risikerer samfunnsvitenskapene å miste monopolet til å analysere samfunnet etter som Big Data-analyser vil erstatte tradisjonelle kvantitative og kvalitative metoder. Med Big Data forsvinner behovet for å trekke utvalg og for å designe case-studier. Med Big Data er N = alle.

For Savage og Burrows (2007) hviler de empiriske samfunnsvitenskapene på distinkte empiriske redskaper, survey og det kvalitative forskningsintervju, som i stadig mindre grad blir sentrale innenfor den forskningsinfrastrukturen som kunnskapskapitalismen utgjør. Samfunnsforskningens metoder har en historisitet: deres sentralitet er knyttet til den øvrige kunnskapsproduksjonen dvs. forhold som kjennetegner både andre kunnskapsprodusenter og kunnskapsbehovene i samfunnet. Surveyens suksess skyldes de mulighetene den gir til generalisering og prediksjon ved hjelp av begrensede ressurser. Den var svært kostnadseffektiv da den kom, men er i økende grad utfordret av fallende responsrater og av konkurranse fra kommersielle aktører (markedsanalyse). Gitt tilgangen til digitale transaksjonsdata fremstår Big Data som mer effektive – fordi slike data er komplette og baserer seg på reelle transaksjoner.

Big Data gjør det også mulig å analysere flere sosiale objekter med kvantitative metoder: tekst, bilder, videoer osv. Savage og Burrows mener det kvalitative intervjuet er mindre egnet til å generere sofistikerte forståelser av de veldig ulike og varierte verdensanskuelser som eksisterer i dagens samfunn. Det kvalitative intervjuet risikerer i økende grad å bli erstattet av metoder (basert på Big Data og web-mining)[2] som vil gi mulighet til å analysere kvalitative objekter i stor skala samt til å generalisere funnene til hele populasjoner.

Big Data vil kunne utfordre samfunnsforskningens tradisjonelle datagrunnlag på flere måter. For det første utgjør Big Data en ny kilde for datainnsamling som fanger opp både handlinger (transaksjonsdata) og meninger (publiserte tekster og bilder), ikke bare holdninger og selvrapporterte handlinger. For det andre muliggjør Big Data innsamling av alle relevante data og ikke bare et utvalg, noe som kan oppfattes som mer pålitelig enn generalisering basert på et utvalg. For det tredje kan prediksjon basert på Big Data komme til å erstatte kausale analyser basert på «Small Data».

Big Data genererer både utopiske og dystopiske profetier om hvordan de vil kunne påvirke både samfunnsutvikling og samfunnsforskning (boyd & Crawford 2012). Teknologien produserer ofte både positive og negative effekter. Big Data vil etter min mening forandre samfunnsforskningen. Det betyr ikke at de tradisjonelle samfunnsforskningsmetodene vil bli utdaterte og forsvinne, men det innebærer at nye data og nye metoder vil være tilgjengelig og vil kunne anvendes på problemstillinger som er sentrale for samfunnsforskning. Spørsmålet er om disse metoder og data vil bli benyttet av samfunnsforskerne, eller om deler av samfunnsforskningen vil konstituere et nytt felt for den voksende computervitenskapen.

BIG DATA OG PERSONVERN

Bruk av sosiale medier har ført til en forvitring av skillet mellom det offentlige og det private som utfordrer personvernet. Både den teknologiske utviklingen og de nye sosiale praksisene som utvikler seg i samspill med de teknologiske verktøyene, bidrar til å endre vårt forhold til hva som er privat og til personvernet. Samtidig mister vi i økende grad kontroll over vår personlige informasjon (poster på sosiale medier, søkemotor-historie, telefontrafikk, innkjøp på nettet, GPS-data osv.).

På tross av at personvern – «retten til å være for seg selv» – er et grunnleggende trekk ved det å være et menneske i det moderne samfunnet, er både definisjonen og forståelsen av begrepet et omstridt tema innenfor både filosofi (Nissenbaum 2010) og psykologi (Joinson & Paine 2009). Psykologene er enige om at personvernet er komplekst og flerdimensjonalt. De er derimot uenige om hvilke dimensjoner (fysisk, interaksjon, informasjon, tilgjengelighet, ekspressiv) som er best egnet for å definere personvernet. Når det gjelder spørsmålet om hvordan digitale nettverk påvirker personvern, mener vi det er informasjonsdimensjonen og den ekspressive dimensjonen som er de meste relevante.

Informasjonsdimensjonen ved personvernet gjelder individets rett til å bestemme og kontrollere hvilken informasjon som skal formidles til hvem. Den ekspressive dimensjonen dreier seg om retten til å beskytte en sfære hvor individet kan uttrykke seg uten å være underkastet myndighetenes eller andres sosiale press og kontroll. Teknologien bak digitale nettverk medfører nye teknologiske muligheter både når det gjelder sporing og overvåking og aggregering og analyse av digital informasjon (Nissenbaum 2010). Disse teknologiske mulighetene påvirker hvordan informasjon om hvert enkelt individ sirkulerer og kan bli (mis)brukt, og hvordan denne bruken berører og i noen tilfeller truer personvernet til den enkelte.

Digitalisering har ført til teknologisk utvikling ikke bare når det gjelder kommunikasjon, men også i form av en rekke verktøy som kan brukes for å fange opp og digitalisere informasjon (digitalt foto-

grafi, lydopptak), samt digitale nettverksoverføringsmekanismer som gjør det mulig å fange opp og overvåke kommunikasjon. Kombinert med veksten i datalagringskapasitet og databeregningskraft, fører disse teknologiske verktøyene til utvidede muligheter for sporing og overvåking av individenes handlinger og kommunikasjon.

Big Data bidrar også til å utfordre personvernet. De fleste selskapene som tilbyr webbaserte tjenester (sosiale medieplattformer som Facebook eller Twitter, søkemotorer som Google eller Bing og dataprogramleverandører som Microsoft eller Apple), lagrer kontinuerlig data om hver enkelt brukers profil, sosiale graf (nettverk av «venner» og følgere på sosiale medier) og webtrafikk. Disse databasene utgjør en enorm og rik mengde informasjon som kan analyseres ved hjelp av «datamining»-metoder. Dette er teknikker for å lage personlige brukerprofiler, som i neste omgang kan anvendes for å tilby målrettet reklame og markedsføring, eller for å tilby produkter gjennom anbefalingssystemer (for eksempel Amazon.com). I kjølvannet av denne utviklingen har en ny nisje oppstått for informasjonsmeglere som selger informasjon fra webtrafikk til en rekke private og offentlige aktører. Ved å benytte «metadata» for å koble sammen datakilder på individnivå, gir Big Data mulighet til å samle og analysere omfattende og detaljert informasjon om en persons liv, aktiviteter, preferanser og ytringer. Begrensede digitale spor som blir lagt enkeltvis på ulike webtjenester eller gjennom apparater som er koblet til Internett, kan utfordre retten til privatliv når de blir aggregert og gjennomanalysert ved hjelp av Big Data-teknologier.

Datamining og maskinlæringsteknologier kombinert med Big Data danner i økende grad en trussel mot ytringsfrihet og personvern. Både regjeringer og private selskaper kan overvåke og analysere kommunikasjon som foregår digitalt. Aggregering av data på tvers av ulike brukerkontoer (for eksempel Google gmail, YouTube, Chrome, Google+ osv.) øker muligheten til å samle omfattende mengder av informasjon om en persons liv, vaner, preferanser, handlinger og meninger. Individuell kontroll over og samfunnsregulering av personlig informasjon er vanskelig å utøve ved hjelp av nasjonal lovgivning fordi dataene som er tilgjengelig digitalt i økende grad er kontrollert av globale selskaper og er i privat eie (Facebook, Google osv.) der brukerne har gitt fra seg rettighetene sine for å kunne benytte seg av tjenestene.

Mark Zuckerberg, Facebooks gründer, fikk mye kritikk i 2010 etter å ha uttalt at personvern ikke lenger er en sosial norm. Digitale mediers utvikling og, nylig, lekkasjene om NSAs overvåkingsprogram PRISM har bare ytterligere bekreftet Zuckerbergs påstand. Personvern er blitt et av de sentrale samfunnsproblemene knyttet til den «digitale alderen». I boken *The New Digital Age* (2013) skisserer Eric Schmidt (Google CEO) og Jared Cohen (Google Idea-direktør) et fremtidig senario for endringene den digitale revolusjonen vil medføre. På tross av at boken er preget av teknologioptimisme, understreker forfatterne hvor sentralt personvern og identitetsbeskyttelse vil bli i fremtidens digitaliserte samfunn. Med den voldsomme økningen i mengden lagrede data mener forfatterne det er stor risiko for at individer blir fratatt kontroll over personlig informasjon i det digitale rommet. Risiko for uautorisert tilgang, manipulering og stjeling av online-identiteter vil også øke. Personlig identitet og personlige data vil ifølge for-

fatterne bli det mest verdifulle godet i den digitale alderen. Med videreutviklingen av Big Data blir utfordringen for medborgere, regjeringer og private selskaper å forutsi hvilke tiltak som vil være hensiktsmessige å gjennomføre for å gjenvinne kontroll over vår private informasjon og identitet online.

I dag er utviklingstrenden kjennetegnet av en ideologi og en digital forretningsmodell som bygger på en byttehandel hvor gratistjenester utveksles mot privat informasjon. Personlig informasjon blir en markedsvare som online-brukere gir bort for å få tilgang til gratis digitale tjenester og som internettselskaper selger videre til reklameannonsører for å kunne finansiere sine online-tjenester. I boken *To Save Everything Click Here* kritiserer Evgeny Morozov (2013) den nye ideologien som kjennetegner internettselskaper lokalisert i *Silicon Valley*. Denne ideologien er ifølge Morozov drevet av viljen til å forbedre nesten alt fra politikk til kokkekunst. Ut fra dette perspektivet er internettselskaper som Google, Amazon, Facebook og Apple drevet av en tro på datateknologier som løsning på de fleste menneskelige problemer. Gjennom algoritmer, databaser og digitale nettverk skal verden kunne effektiviseres og forbedres. Denne ideologien er drivkraften bak en rekke banebrytende innovasjoner fra e-bøker til sosiale medier. Den er også drivkraften bak utviklingen av en ny forretningsmodell hvor personlig informasjon er en ressurs. Personlig informasjon har en markedsverdi fordi den danner grunnlaget for reklameinntekter eller bidrar til å gjøre tjenestene mer effektive og personalisert. Bruk av anbefalingssystemer som bygger på maskinlæringsalgoritmer anvendt på store mengder av personlig informasjon, er kjernen i den nye forretningsmodellen.

Big Data har allerede bidratt til at våre sosiale normer relatert til personvern har endret seg. Bruk av sosiale medier innebærer ofte en større personlig eksponering og formidling av privat informasjon i det offentlige eller kvasi-offentlige rommet. Normendringene blir forsterket av den teknologiske utviklingen. Jo større andel av vår aktivitet som involverer elektroniske hjelpemidler koblet til Internett, desto større blir mengden av personlig informasjon som er samlet, lagret og som blir behandlet for ulike formål. Denne utviklingen skaper en utfordring på samfunnsnivå når det gjelder beskyttelsen av privatsfæren og for personvernet. For samfunnsforskningen er Big Data tveegget: på den ene siden åpner tilgjengeligheten av nye data for nye muligheter; på den andre siden skaper denne utviklingen nye etiske problemstillinger som ikke lar seg løse med et enkelt regelverk.

SAMFUNNSFORSKNINGENS REGULERING I BIG DATA-ALDEREN

Tilgangen til nye digitale data som kan benyttes for samfunnsforskningsformål har skapt en gråsone for samfunnsforskningen. Mengder av nye data er tilgjengelige for analyse, men dagens regelverk begrenser kraftig bruken av disse dataene for forskning samtidig som dataene kan benyttes fritt for kommersielle formål. Min påstand er at samfunnsforskningen nå er konfrontert med et anakronistisk regelverk som ble designet for Small Data og som ikke er tilpasset dagens teknologi i en globalisert verden.

I Norge og i Europa for øvrig er webdata underlagt de samme retningslinjer som andre typer data. Med hjemmel i menneskerettighetskonvensjonen har personopplysningsloven som hensikt å

verne privatpersoner mot krenkelser og mot bruk av bilder eller personopplysninger uten samtykke. Ifølge personopplysningsloven krever elektronisk lagring av personlig informasjon (også når denne informasjonen har blitt offentliggjort) tillatelse fra hver enkelt person. Ifølge loven er personopplysninger en opplysning eller vurdering som kan knyttes til et individ som enkeltperson. Når en virksomhet behandler personopplysninger, skal dette i størst mulig grad være basert på et *samtykke*. Dersom en virksomhet behandler personopplysninger uten at den har innhentet samtykke, må den ha et annet *rettslig grunnlag* (som er tilfellet for eksempel når det gjelder ileggelse av skatt eller utbetaling av trygd).

Dette reiser spørsmål om hvilken status offentlig tilgjengelige data i sosiale medier, som for eksempel Twitter, skal ha i forskningssammenheng. Er det for eksempel rimelig at individene som blir forskningsobjekter (bloggere, twitterbrukere osv.) skal måtte samtykke til bruk av sine offentlig tilgjengelige data i forskning?

Dagens regelverk betyr i praksis at bruk av Big Data som inneholder personopplysninger til samfunnsforskningsformål krever at forskerne innhenter samtykke for millioner av individer. For eksempel vil et forskningsprosjekt som vil analysere bruk av Twitter i valgkampen måtte spørre millioner av brukere om samtykke til å lagre deres tweets, mens de allerede er lagret og offentlig tilgjengelig gjennom Twitter API.

Regelverket som ble utformet i en tid hvor «Small Data» ble lagret i strukturerte databaser med en bestemt eier, har i dag flere uheldige konsekvenser for anvendelse av Big Data til samfunnsforskningsformål.

For det første er dette strenge kontrollregimet når det gjelder bruk av webdata til forskningsformål hemmende for forskning og i praksis lite effektivt i en globalisert og kommersialisert verden. Globalisering medfører at det stedbundne regelverket ikke lar seg håndheve eller skaper gråsoner for hva som er tillatt. I hvilken grad, for eksempel, er lagring av norske webdata i Amazon Sky i USA underlagt det norske regelverket? Kommersialisering medfører at norske personlige data kan kjøpes på utenlandske markeder hos *data brokers* eller gjennom datatjenester som har spesialisert seg på å samle og selge denne typen data. Private internettaktører samler og selger mengder av personlige data uten å bryte loven, ved å tilby gratistjenester hvor tilgang forutsetter at brukeren godkjenner en rekke vilkår (som de fleste brukere ikke er klar over) og blant annet gir fra seg rettigheter til kontroll av personlig informasjon. En sentral utfordring med Big Data er at anonymisering av dataene i økende grad er en illusjon. Metadata som er knyttet til transaksjonsdata (for eksempel e-post adresse eller IP-adresse), kan brukes for å koble ulike datakilder sammen og gjør personlig identifikasjon enkelt.

For det andre: Big Data er i økende grad i privat eie og ikke tilgjengelig for forskningssamfunnet. Mengder av data samles inn, lagres og analyseres rutinemessig, med mer og mer sofistikerte metoder. Disse er i privat eie, og bruken er ikke underlagt de samme restriksjonene som forskere må underlegge seg. Denne utviklingen samt forskjeller i personvernlovgivningen mellom Europa og resten av verden har negative konsekvenser for samfunnsforskningens bruk av Big Data. For det første fører dette til et økende *digitalt skille* mellom europeisk samfunnsforskning på den ene siden og private selskaper og amerikanske forsk-

ningsmiljøer på den andre. Internettselskaper som Facebook, Twitter, og Google og utvalgte universiteter som samarbeider med disse selskapene (Stanford, MIT osv.), har eksklusiv tilgang til store mengder sosiale data som de fleste forskere ikke har tilgang til. Forskerne knyttet til disse miljøene kan produsere forskning som andre forskere ikke kan, samtidig som denne typen forskning ikke lar seg reprodusere eller evaluere siden dataene er privatisert. På sikt kan dette føre til en privatisering av samfunnsforskningen. Private internettselskaper har, i motsetning til offentlig finansiert forskning, ingen plikt eller ansvar for å gjøre sine data tilgjengelig for forskningsfellesskapet eller plikt til å utsette sine funn for det vitenskapelige fellesskapets kritikk.

For det tredje viser personvern gjennom nasjonale regelverk seg å være lite effektivt. Det rammer hovedsakelig forskning og i liten grad kommersiell bruk av Big Data. Det er flere grunner til dette. For det første er personvernregulering mer liberal i USA og i mange andre land enn den er i Europa. For det andre: hvis brukere formelt sett må samtykke til datalagring for å kunne bruke kommersielle tjenester og applikasjoner (som Facebook, Google osv.), viser det seg at de fleste ikke leser vilkårene for tjenestebruk og egentlig ikke er klar over hva de har samtykket til. For det tredje: mye personlig data fra sosiale medier og nettbruk er kommersielle produkter som selges av spesialiserte selskaper (data brokers) utenfor Europa.

Å ivareta personvern i en Big Data-alder krever derfor nye typer virkemidler. Gitt den globale og desentralisert arkitekturen som kjennetegner Internett, blir en geografisk og stedbundet statlig regulering av Internett stadig utfordret av ny teknologisk utvikling som gjør det mulig å unngå regulering og som dermed er lite effektivt i et globalt landskap.

I tillegg er regulering ofte innebygget i selve koden som definerer hvordan de ulike lagene som konstituerer Internett fungerer og interagerer, og som er kontrollert av private aktører. Digital kommunikasjon gjennom Internett innebærer ulike nivåer av infrastruktur, protokoller og dataprogrammer for transmisjon, mottak og behandling av digital informasjon. Alle disse elementene er programmert, og programmene (koden) har en regulerende funksjon innebygd i seg i den forstand at programmene bestemmer hva som er mulig og umulig å gjøre på Internett. Internetts regulering skjer først og fremst gjennom programmene som konstituerer Internett.

Statlig regulering kan også ha utilsiktede og uønskede effekter i form av informasjonsovervåking, kontroll, sensur og svekket ytringsfrihet. Selvregulering eller markedsbasert regulering av Internett er heller ikke samfunnseffektivt. Markedsbasert selvregulering har en tendens til ikke å ta hensyn til teknologiens sosiale og demokratisk uønskede effekter. Nettverkseffekter[3] som kjennetegner internettbaserte industrier fører til konsentrasjon i informasjonsindustrien og svekkelse av konkurranse og forbrukernes makt. Brown og Marsden (2013) foreslår en flerinteressent (*multistakeholder*) styringsmodell (*governance*) av Internett hvor både stater, sivilsamfunn og markedsaktører forhandler om hvordan Internett skal reguleres. I en slik modell vil både sivilsamfunnsaktører og stater kunne bidra til å definere, sammen med industrielle aktører, Internetts tekniske standarder. Programmeringskoden som muliggjør Internetts infrastruktur og digitale applikasjoner har en regulerende

funksjon (Lessig 2006) og er dermed viktigere enn regelverket. Koden definerer hvilke handlingsmuligheter (*affordances*) en teknisk protokoll eller applikasjon tillater eller ikke. For sivilsamfunns- og statlige aktører er det viktigere og mer effektivt å kunne påvirke de tekniske løsningene (koden) enn å vedta et lovverk som ikke lar seg håndheve. En slik styringsmodell som regulerer Internett også gjennom koden og ikke bare gjennom loven, vil kunne ivareta, i tillegg til industriens tekniske behov og økonomiske interesser, statenes og medborgernes interesser samt demokratiske og samfunnsmessige hensyn, inkludert personvern.

I påvente av reguleringsmekanismer tilpasset det globale Internett, er det behov for å tilpasse personvernreglene som gjelder for samfunnsforskning i Big Data-alderen. En mulig løsning vil være mindre lovbasert regulering og mer personlig etisk ansvar for forskerne. Internasjonalisering av datainnsamling, datalagring og behandling, kommersialisering av personlig data, *cloud computing* og internasjonal publisering medfører at nasjonale regler ikke fungerer effektivt og bidrar til å straffe nasjonale og europeiske forskningsmiljøer i den internasjonale konkurransen. Individuell godkjenning av hvert enkelt individ når det gjelder elektronisk datalagring av personlig informasjon, er ikke tilpasset den nye virkeligheten hvor Big Data består av informasjon fra millioner av individer. Konsesjon for datainnsamling, lagring og analyse av Big Data (inkludert webdata) burde gis til forskere og forskningsmiljøer som forplikter seg til å overholde bestemte etiske og systemiske regler, både når det gjelder lagringssikkerhet, analysemetoder og publisering. Dette må være etiske regler som garanterer individenes anonymitet og beskyttelse av personlig opplysninger uten individuell forhåndsgodkjenning. Individuelt samtykke for hver enkelt bit av data som er tilgjengelig på Internett er ikke tilpasset de nye mulighetene som åpner seg for samfunnsforskning med Big Data. Det er likevel viktig å ivareta etiske hensyn samt å garantere at disse nye mulighetene ikke skal føre til misbruk. Det er på tide at både forskningsmiljøer og myndighetene tar utviklingen knyttet til Big Data på alvor, for å finne løsninger som garanterer både at nye forskningsdata vil kunne bli benyttet og at medborgerne blir beskyttet mot ulike former for misbruk av disse dataene.

Noter

1 Distribuert databehandling har blitt muliggjort gjennom to innovasjoner: den ene er Google File System sin *open source*-versjon Hadoop som styrer lagring og beregning gjennom alle datamaskiner som danner en *cluster*, og den andre er algoritmen «Map Reduce» som effektiviserer databehandling av store mengder data. Parallelt har også nye databaseløsninger blitt utviklet som har erstattet det dominerende databaseparadigmet kjent som SQL (Structured Query Language) for Big Data applikasjoner. Det nye databaseparadigmet, relasjonell eller noSQL database, er tilpasset interaktive webteknologier og distribuerte databehandlingssystemer.

2 Web mining er anvendelsen av data mining-teknikker i analyser av Internett. Med data mining menes en automatisk eller delvis automatisk analyse

av store mengder digitale data med tanke på å finne mønstre. Avhengig av målet for analysen kan web mining deles inn i tre forskjellige typer: webbruk-mining (mønstre i hvordan Internett brukes – for eksempel hvor mange som søker på influensasymptomer på et gitt tidspunkt), webinnhold-mining (innhenting av data og informasjon fra websidenes innhold) og webstrukturmining (analyse av nettverksstruktur på Internett – for eksempel hvordan følgere av de ulike politiske partiene i Norge er koblet i nettverk).

3 Produkter fra nettverksbaserte industrier (som strømleveranse, telefoni, internettkommunikasjon) har bestemte egenskaper som komplementaritet, eksterne effekter, stordriftsfordeler, høye byttekostnader og *lock-in*, noe som bidrar til at markedskonkurranse ikke fungerer effektivt og at disse industriene nærmer seg en monopolistisk posisjon. (Se for mer informasjon: Shy 2001.)

Referanser

Boyd, d. & K. Crawford (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5):662–679.

Brown, I. & C. T. Marsden (2013). *Regulating Code. Good Governance and Better Regulation in the Information Age*. Cambridge: MIT Press.

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. New York: Springer.

Joinson, A. N. & C. B. Paine (2007). Self-Disclosure, Privacy and the Internet. I A. N. Joinson, K. McKenna, T. Postmes & U. Reips (red.), *Oxford Handbook of Internet Psychology*. Oxford: Oxford University Press.

Lessig, L. (2006). *Code 2.0*. New York: Basic Books.

Mayer-Schönberger, V. & K. Cukier (2013). *Big Data. A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt.

Morozov, E. (2013). *To Save Everything Click Here. The Folly of Technological Solutionism*. New York: Public Affairs.

Nissenbaum, H. (2010). *Privacy in Context. Technology, Policy and the Integrity of Social Life*. Standford: Standford University Press.

O'Reilly, T. (2005). *What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software*. Hentet fra http://oreilly.com/web2/archive/what-is-web-20.html

Savage, M. & M. Burrows (2007). The Coming Crisis of Empirical Sociology. *Sociology*, *41*(5):885–899.

Schmidt, E. & J. Cohen (2013). *The New Digital Age. Reshaping the Future of People, Nations and Business*. New York: Knopf.

Shy, O. (2001). *The Economics of Network Industries*. Cambridge: Cambridge University Press.

Siegel, E. (2013). *Predictive Analytics*. Hoboken: Wiley.