DIGITALA SYSTEM
SYSTEMS ENGINEERING

Transparenta algoritmer i försäkringsbranschen: Transparens och hållbarhetsprediktion

Ulrik Franke

Bilaga till RISE Rapport 2021:04

Transparens och hållbarhetsprediktion

Denna bilaga utgör slutredovisning för det avslutande delprojektet *Transparens och hållbarhetsprediktion* inom projektet *Transparenta algoritmer i försäkringsbranschen* (TALFÖR, P4/18) finansierat av Länsförsäkringars forskningsfond.[1]

Tyngdpunkten i delprojektet har legat på samarbete med projektet *Hållbarhetsprediktion med maskininlärning*, även känt som *Hållbarhetsrobot* (P8/18) likaledes finansierat av Länsförsäkringars forskningsfond.[2] De två projekten löpte parallellt under fyra månader; februari—maj 2021, när Hållbarhetsrobotprojektet avslutades. Därefter löpte TALFÖR-delprojektet ytterligare fem månader; juni—oktober. Utöver samarbetet med Hållbarhetsrobotprojektet har det inom ramen för TALFÖR-delprojektet genomförts ytterligare några aktiviteter relaterade till bredare transparensfrågor.

Transparens, investeringar och cybersäkerhet

Den första huvudpunkten i samarbetet med Hållbarhetsrobotprojektet har varit en studie av transparens, investeringar och cybersäkerhet genomförd av Ulrik Franke (projektledare för TALFÖR) tillsammans med Jan Svanberg (projektledare för Hållbarhetsrobotprojektet). Utgångspunkten är att det finns likheter mellan å ena sidan utsläpp och å andra sidan dålig cybersäkerhet. Båda är vad nationalekonomer brukar kalla för *negativa externaliteter*; handlingar där den som handlar inte bär hela kostnaden. Konkret bär en enskild förorenare inte hela kostnaden för nedsmutsningen och investerar därför inte tillräckligt i reningsåtgärder. På samma sätt bär en enskild verksamhet med dålig cybersäkerhet inte hela kostnaden för detta: också andra drabbas när skadlig kod sprids, det uppstår avbrott i tjänster, eller personuppgifter läcker. Därför anses det allmänt i litteraturen att det investeras för lite i cybersäkerhet.

En åtgärd för att komma till rätta med denna incitamentsproblematik som flitigt har diskuterats i litteraturen är att *offentliggöra* cyberincidenter. Om incidenterna blir kända så kan kunder, leverantörer, investerare med flera straffa dålig cybersäkerhet genom att välja att sluta köpa eller sälja ett företags produkter, sälja av dess aktier etc. Därmed uppstår ett ökat incitament att undvika cyberincidenter, alltså att höja sin cybersäkerhetsnivå.

Det unika bidraget i den studie som har genomförts inom projektet är att analysera cybersäkerhetsproblemet genom en systematisk jämförelse med utsläppsproblemet. Mer precist jämförs de tänkbara men ännu inte särskilt empiriskt välstuderade effekterna av offentliggörande av cyberincidenter med de mycket mer empiriskt välstuderade effekterna av offentliggörande av koldioxidutsläpp. Jämförelsen listar fem likheter och fem skillnader mellan de två typerna av offentliggörande. Utifrån detta finner studien att ökad redovisning av cyberincidenter eller de omständigheter som orsakar dem skulle kunna leda till ökade kostnader för eget kapital och skuld för företag med många eller allvarliga cyberincidenter samt även till aktieägaraktivism och eventuellt minskande efterfrågan. Däremot bedöms dessa effekter troligen vara *mindre* för offentliggörande av

[1] Fullständig slutrapport finns tillgänglig i DiVA.
[2] Fullständig slutrapport finns tillgänglig hos forskningsfonden.

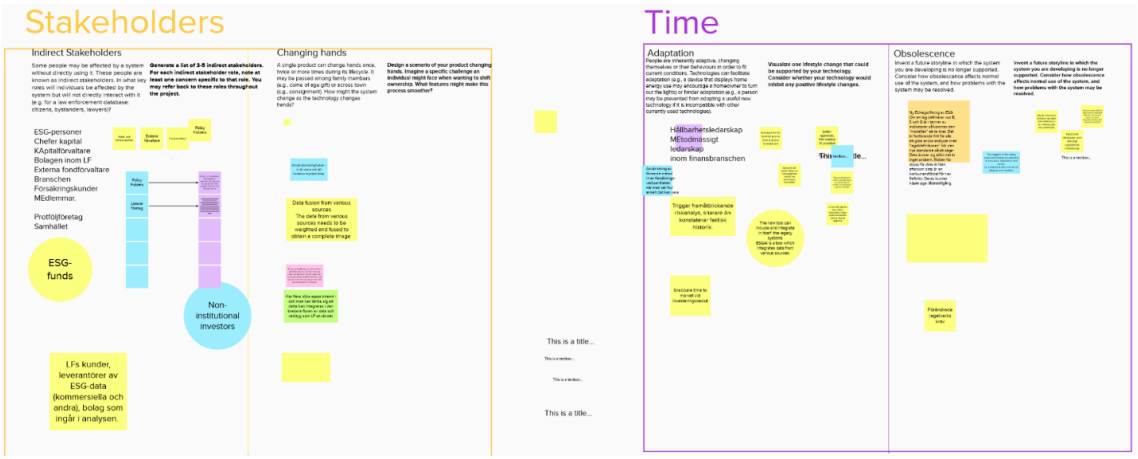
cyberincidenter jämfört med motsvarande effekter för offentliggörande av koldioxidutsläpp.

Den samförfattade artikeln som beskriver studien (U. Franke & J. Svanberg. *Cyber incident disclosure—lessons from CO_2 disclosure*) genomgår just nu granskning hos en vetenskaplig tidskrift.

Utvecklingen av hållbarhetsroboten

En andra huvudpunkt i samarbetet med Hållbarhetsrobotprojektet har bestått i att tillgängliggöra den kompetens som har byggts upp i TALFÖR-projektet för att göra hållbarhetsroboten så användarvänlig och förklarbar som möjligt. Projektmedlemmar från TALFÖR har därför deltagit i ett antal (distans-)möten mellan Hållbarhets-robotprojektet och mottagarna på Länsförsäkringar under mars och april. Genom att på detta sätt lyssna av aktuella behov och frågeställningar och tillföra ett användar- och förklarbarhetsperspektiv har TALFÖR gjort ett litet bidrag till en god leverans. Inom ramen för den korta tiden till leverans bedömdes det däremot inte finnas tid till mer systematisk utvärdering av olika lösningar i form av exempelvis experiment.

Den 27 april ledde TALFÖR en *Value Sensitive Design*-workshop med såväl projektmedlemmar från Hållbarhetsrobotprojektet som mottagare på Länsförsäkringar. Workshopen följde snarlik metodik som tidigare har använts och publicerats inom TALFÖR (J. Dexe et al. Towards Increased Transparency with Value Sensitive Design. I: *Artificial Intelligence in HCI. HCI International 2020.*, ss 3–15. Springer, juli 2020).



Figur: Dokumentation av workshopen.

Fyra fenomen relaterade till hållbarhetsroboten diskuterades ingående:

Indirekta intressenter (eng. *indirect stakeholders*): Workshopen gjorde det tydligt att det finns väldigt många intressenter som (potentiellt) påverkas av en hållbarhetsrobot. Några av de identifierade är ESG-experter, kapitalförvaltare, externa fondförvaltare, branschen i stort, försäkringskunder, portföljföretag och leverantörer av ESG-data. Det är rimligt att tro att de alla påverkas på lite olika sätt av olika förklarbarhetsstrategier, men workshopen gav inga slutgiltiga svar.

Byta händer (eng. *changing hands*): När verktyget lämnas över från forskarna till användarna blir förklarbarhet viktig. Förstår alla vad hållbarhet är? Blir det svårt att förstå hur data från olika källor fusioneras? Och vad händer om konkurrenter läser

pressmeddelanden och akademiska publikationer så att de kan bygga liknande hållbarhetsrobotar själva?

Beteendeförändring (eng. *adaptation*): Det här temat gav upphov till många spännande tankar med förhoppningar om ett nytt hållbarhetsledarskap i finansbranschen, framåtblickande riskanalys istället för historiebundenhet, ny förståelse hos bolagen för vikten av förebyggande arbete och påverkan på lagstiftaren. Möjligen var deltagarna lite väl låsta i att föreställa sig positiva konsekvenser – det är också viktigt att leta efter negativa konsekvenser så att de kan förebyggas i tid.

Tiden springer ifrån produkten (eng. *obsolescence*): Frågeställningen om åldrande ledde framförallt till reflektioner kring dataleverantörernas och lagstiftarens roll för att sätta *de jure*- och *de facto*-standarder. Det gjordes också mer tekniknära iakttagelser kring att vissa bedömningar lever för länge och att modellen kan behöva förses med en explicit funktion för att glömma alltför gammal data.

Workshopen togs emot positivt av deltagarna även om långt ifrån alla frågeställningar och iakttagelser enkelt gick att åtgärda inom ramen för utvecklingsprojektet.

Upphandling av transparenta och förklarbara tjänster

De flesta verksamheter har inte den AI-expertis som krävs för att på egen hand utveckla framtidens smarta automatiserade lösningar inom områden som bildigenkänning, texthantering eller riskbedömning. Istället för att uppfinna hjulet på nytt upphandlas färdiga komponenter från andra. Dessa komponenter kan sedan sättas samman och anpassas för att på bästa sätt passa den aktuella verksamheten. Detta är i grunden av godo, eftersom det möjliggör för många fler organisationer att ta del av de potentiella fördelarna med modern AI-teknik.

Samtidigt finns det nackdelar. Den som inte själv utvecklar sina AI-lösningar riskerar att hamna i beroendeställning till leverantören och köpa en svart låda. Det kan i så fall leda till ett obehagligt uppvaknande om produkten exempelvis visar sig diskriminera mot vissa grupper. Det finns gott om exempel på sådana fall, oftast för att olika grupper inte representerats ordentligt i träningsdata från början. Att kunna dra nytta av fördelarna utan att drabbas av nackdelarna med AI som köps in utifrån kan ur detta perspektiv ses som en nyckelförmåga hos en modern inköpsfunktion.

Utifrån dessa frågeställningar utforskade delprojektet tillsammans med Länsförsäkringars inköpsfunktion möjligheterna att genomföra dels förmågehöjande utbildningsverksamhet, dels forskningsverksamhet tillsammans, exempelvis genom att söka forskningsfondens utlysning om interna AI-projekt. Dialogen pågick under februari—april (9 februari, 16 februari, 19 februari, 1 mars, 22 mars) och kulminerade i ett seminarium den 14 april där forskningsfonden tillsammans med projektet gav inköpsfunktionen en grundläggande genomgång av AI och dess påverkan på Länsförsäkringar.

Dialogen ledde till ett gemensamt beslut att i detta skede inte söka forskningsfondens utlysning om interna AI-projekt. Däremot noterades ett intresse för frågorna och en möjlighet att ta upp samarbetet i framtiden.

Övriga projektaktiviteter

Delprojektet har under fortsättningsperioden genomfört en internationell empirisk studie av hur GDPR-rätten till förklaring av automatiserat beslutsfattande fungerar i praktiken. Metodmässigt är det en uppskalning av den tidigare studie som genomfördes inom TALFÖR 2018—2019. I den nu aktuella studien har kunder i försäkringsbolag i Danmark, Finland, Nederländerna och Polen bett om förklaringar på hur premierna i deras hemförsäkringar sätts utifrån GDPR, paragraf 15. Totalt rör det sig om förfrågningar till 26 olika försäkringsbolag, inklusive de sju svenska som redan undersökts. Ett konkret resultat är att försäkringsbranschen tycks vara mer tillmötesgående än andra branscher. Ett annat resultat är att lagen är svårtolkad, bland annat eftersom de olika språkversionerna av GDPR kan tolkas på lite olika sätt. Tidskriftsartikeln som beskriver studien har reviderats efter en första granskningsomgång och är nu under granskning igen. Ytterligare en fortsättningsstudie med liknande GDPR-förfrågningar till ett bredare urval av företag, bortom försäkringsbranschen, pågår i skrivande stund och kommer inom kort att skickas in till en vetenskaplig tidskrift för granskning.

Delprojektet har också studerat filosofiska aspekter av transparens och förklarbarhet i automatiserat beslutsfattande. I en publicerad tidskriftsartikel (U. Franke. Rawls's Original Position and Algorithmic Fairness, *Philosophy & Technology*, 2021) diskuteras John Rawls ursprungsposition och i vilken utsträckning den kan tillämpas för att säkerställa rättvisa algoritmer. Ursprungspositionen är ett mycket omskrivet tankeexperiment, där invånarna i ett samhälle antas samlas bakom en ”okunnighetens slöja” för att bestämma samhällets grundstruktur. Okunnighetens slöja innebär att ingen vet sin egen position i samhället, sin egen begåvning eller någon annan av sina egna egenskaper. Tanken är att denna okunskap ska leda till mer opartiska beslut i allas intresse, snarare än snävt egenintresse. Idén kommer alltså från politisk filosofi, men har på senare år ofta använts i AI-sammanhang. Den publicerade artikeln visar att det finns stora och relevanta skillnader mellan ursprungspositionen såsom den används i bredare politisk filosofi och en motsvarande ursprungsposition för algoritmrättvisa. För det första finns det skillnader avseende riskbenägenhet: de skäl som Rawls anför för en riskavers inställning finns i mycket mindre utsträckning i algoritmsammanhanget. För det andra finns det i algoritmsammanhanget ett gränsdragningsproblem som inte finns i det bredare politisk-filosofiska sammanhanget avseende vilka som ska representeras i ursprungspositionen. För det tredje är det i algoritmsammanhanget svårare att definiera dem som har det sämst ställt (en nyckelgrupp i Rawls teori som behöver identifieras för att den s.k. differensprincipen ska gå att tillämpa). För det fjärde så tycks frågeställningarna kring algoritmrättvisa kräva mer kunskap om sannolikheter än vad som tillåts bakom okunnighetens slöja. Ytterligare ett par artiklar med filosofiskt innehåll genomgår i skrivande stund vetenskaplig granskning.

Inom ramen för delprojektet har även en andra omgång av doktorandkursen Transparens i tekniska och sociala system genomförts på KTH under vårterminen 2021.