

Manglende tillid til risikovurderings algoritmer

5. semester projekt • Filosofi & Videnskabsteori E2020 • Roskilde Universitet

Vejleder: Jesper Ryberg

Eksamensnummer: V2025073413

Antal tegn: 106.999

01110011 01110000 01100001 01110TILLID010 01100101 01101110 01110100 01100101
0010ALGORITMER0000 01110010 01101001 01110011 01101001 01101011 01101111 01110110
01110101 01110010 01100100 01100101 01110010 01101001 01101110 01100111 01110011 00100000
01100001 01101100 01100111 01101111 01110010 01101001 01110100 01101101 01100101 01110010
0011111101000010 11COMPAS 01110010 00100000 01101101 01100001 01101110 00100000
01100010 01110010 01110101 01100111 01100101 00100000 01101001 01101110 01110100 01110010
01100001 01101110 01110011 01110000 01100001 01110010 01100101 01101110 01110100 01100101
00100000 01110010 0110UNFAIR1001 01110011 01101001 01101011 01101111 01110110 01110101
01LOOMIS 01100100 01100101 01110010 01101001 01101110 01100111 01110011 00100000 01100001
01101100 01100111 01101111 01110010 01101001 01110100 01101101 01100101 01110010
0011111101000010 11111000 01110010 00100000 01101101 01100001 01101110 001RISK00000
0110001ASSESSMENT 01110101 01100111 01100101 00100000 01101001 01101110 01110100
01110010 01100001 01101110 01110011 01110000 01100001 01110010 01100101 01101110 01110100
01101100 01100111 01101111 01110010 01101001 01110100 01101101 01100101 01110010
0011111101000010 11111000 01INQUIRY110010 00100000 01101101 01100001 01101110 00100000
01100010 01110010 01110101 01100111 01100101 00100000 01101001 01101110 01110100
0111DISKRIMINATION0010 01100001 01101110 011MISTILLID10011 01110000 01100001 01110010
01100101 01101110 01110100 01100101 00100000 01110010 01101001 01110011 01101001 01101011
01110011 00100000 01100001 01101100 01100111 01101111 01110010 01101001 01110100 01101101
011001010 11000101 01110010 00INTRANSPARENS 0011 11100111101000010 11111000
011100TRANSPARENS 01101101 01100001 001101 01101110 001000OPACITY000 01100010
01110010 01110101 01100111 01100101 00100000 01101001 01101110 01110100 01110010 01100001
01101110 01110011 01110000 01100001 01110010 01100101 01101110 01110100 01100101 00100000
01100111 01101111 01110010 01101001 01110100 01101101 01100101 01110010 0011111101000010
11111000 01110010 00100000 011OPAQUE01101 01100001 01101110 00100000 01100010 01110010
01110101 01100111 01100101 00100000 01101001 01101110 01110100 01110010 01100001 01101110
01110011 01110000 011OBJECTION00001 01110010 01100101 01101110 011FAIR10100 01100101
00100000 01110010 01101001 01110011 01101001 01101011 01101111 01101110 01110101 01110010
01100100 01100101 01110010 01101001 01101110 01100111 01110011 00100000 01100001 011011001

Helena Sarah Christiansen (65318)

Frida Valles Kirkegaard (65451)

Casper Jensen (66496)

Bjørn Arnel Iisager (65415)

Abstract

This paper argues that one should be aware of the use of risk assessment algorithms in the criminal justice system in regard to ethical concerns. To further investigate this inquiry this paper discusses four objections, against the use of biased opaque risk assessment algorithms in the criminal justice system. One objection argues that opacity in algorithms creates distrust towards the criminal justice system in general. Another objection argues that opacity in algorithms makes it difficult for higher courts to assess how a given case has been judged. The third objection argues that transparency secures a greater experience of fairness in the juridical process. The fourth objection argues that transparency makes it easier to detect flaws in algorithms. It is concluded in this paper, that an opaque risk assessment algorithm, does not necessarily become more legitimate if it becomes more transparent.

Indholdsfortegnelse

1.0 Motivation/Indledning.....	4
2.0 Problemformulering	5
3.0 Forskningsoversigt.....	5
3.1 Julia Angwin, Surya Mattu, Lauren Kirchner Et al.: <i>Machine Bias</i>	5
3.2 Alyssa M. Carlson: The Need for Transparency in the Age of Predictive Sentencing Algorithms	6
3.3 Vincent Chiao: Transparency: Are Judges Better Than Algorithms?	6
3.4 Danielle Kehl, Priscilla Guo, Samuel Kessler: Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing	7
3.5 Cynthia Rudin: Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead	7
3.6 Jesper Ryberg: Sentencing and Algorithmic Transparency	8
3.7 William Seymour: Detecting Bias: Does an Algorithm Have to Be Transparent in Order to Be Fair?	8
4.0 Redegørelse.....	8
4.1 COMPAS-algoritmen og hvordan denne bruges i det amerikanske retssystem.	8
4.2 Rudins argumentation mod brugen af black box ML modeller.....	12
4.3 COMPAS' intransparens.....	14
4.4 Carlsons argumentation for transparens i risikovurderings algoritmer.....	15
4.5 Om straffepoces.....	16
4.6 The Loomis case: Ethiske komplikationer ved COMPAS-algoritmen	17
5.0 Analyse & diskussion af centrale argumenter	20
5.1.0 Første argument: Intransparente algoritmer bør ikke anvendes, da det vil skabe mistillid i befolkningen	20
5.1.1 Diskussion af argumentet.....	21
5.1.2 Første indvending: Transparens af koden er irrelevant, da befolkningen alligevel ikke forstår den.....	21
5.1.3 Anden indvending: Dommere kan også betragtes som intransparente, så hvis intransparente algoritmer skaber mistillid, så må det samme kunne siges om dommere	22
5.1.4 Tredje indvending: Det er ikke korrekt at lukkede algoritmer vil skabe mere mistillid i befolkningen, da vi også benytter lukkede algoritmer andre steder i samfundet og dette skaber ikke mistillid	26
5.1.5 Opsamling	27
5.1.6 Delkonklusion.....	28
5.2.0 Andet argument: Intransparens vanskeliggør en appeldomstols vurdering af en domfældelse, da begrundelserne bag vurderingen er ukendte	28

5.2.1 Diskussion af argumentet.....	28
5.2.2 Første påstand: Transparens øger præcisionen af dommen.....	29
5.2.3 Anden påstand: Der er flere årsager til at intransparente modeller ikke bør anvendes, end at det nødvendigvis går ud over en appeldomstols vurdering af en sag, blandt andet tiltaltes ret til indsigt i begrundelser bag dommen.....	30
5.2.4 Første indvending: Intransparente algoritmer vanskeliggør ikke en appeldomstols vurdering af en domsfældelse, selvom begrundelserne bag er ukendte	31
5.2.5 Opsamling	31
5.2.6 Delkonklusion.....	32
5.3.0 Tredje argument: Transparens skaber oplevelsen af en mere fair behandling hos den tiltalte	33
5.3.1 Diskussion af argumentet.....	33
5.3.2 Første indvending: Transparens i en algoritme skaber ikke oplevelsen af fair behandling, da det netop kan vises at algoritmen ikke er fuldkommen objektiv.....	34
5.3.3 Anden indvending: En intransparent algoritme kan godt skabe en oplevelse af fair behandling.....	34
5.3.4 Tredje indvending: Alt afhænger ikke af algoritmens transparens	36
5.3.5 Opsummering.....	36
5.3.6 Delkonklusion.....	38
5.4.0 Fjerde argument: Transparens gør det lettere at opdage og rette fejl i algoritmen.....	38
5.4.1 Diskussion af argumentet.....	39
5.4.2 Første påstand: Ekspertter skal have adgang til algoritmens kodning.	39
5.4.3 Første indvending: Kodningen er for kompleks, så selv fagpersoner ikke forstår algoritmen.	40
5.4.4 Anden indvending: Risikovurderings algoritmer optimerer sig selv.....	41
5.4.5 Opsummering.....	43
5.4.6 Delkonklusion.....	43
6.0 Konklusion.....	44
7.0 Litteraturliste.....	47
7.1 Bilag.....	49

1.0 Motivation/Indledning

Da vi startede denne rapport, bundede vores interesse i en nysgerrighed for diskriminations-spørgsmålet i det amerikanske retssystem. Her særligt med henblik på racediskrimination af afroamerikanere i kontrast til en forfordeling af den hvides privilegier; herunder fordele ved køn, det at have en særlig hudfarve og social status. Endvidere besad vi ligeledes en grundlæggende interesse inden for udviklingen af Artificial Intelligence, samt medfølgende etiske diskussioner af dette, som i dag opfattes relevante.

Vi i gruppen, studerende fra Roskilde Universitet, fandt frem til artiklen udgivet af ProPublica: *Machine Bias* (2016) af Angwin, Julia, Surya Mattu, Lauren Kirchner Et al., som for alvor fik tændt ild i vores passion for en god diskussion om, hvorvidt vi bør bruge risikovurderings algoritmer. I denne artikel beskrives en bekymring for, og problematik af, brugen af COMPAS-algoritmen i det amerikanske retssystem. ProPublica påpeger at denne algoritme, som bruges til vurderinger af en tiltaltes sandsynlighed for tilbagefald til kriminalitet, anses for at være gennemsyret af biased diskriminerende forudindtagelser, som ikke er tilgængelige eller mulige at rette op på, da COMPAS er en intransparent algoritme.

Efter en gennemlæsning af denne artikel, blev vores interesse forstærket og førte os til et valg om, at tage udgangspunkt i den forudgående og stadigvæk aktuelle debat. I stedet for at fokusere på spørgsmålet om diskrimination i algoritmerne, endte vi i en diskussion om, hvorvidt det at algoritmer er intransparente, såsom COMPAS-algoritmen, faktisk er problematisk.

Efter at have tilegnet os en mere omfattende viden indenfor denne debat, har vi set argumenter både for og imod den etiske forsvarlighed, i brugen af sådan en algoritme, til risikovurdering i samarbejde med domsafsigelser i juridiske sammenhænge.

Overvejelser om, hvorvidt dette er forsvarligt, er i vores øjne en nødvendighed, da kodning og forståelse for algoritmens virken ikke er tilgængelig for offentligheden. Disse algoritmer bruges i dag til domfældelse og har indflydelse på en dommers vurdering af en given sag - da dette går ind og kan have direkte indflydelse på et individs liv, er en stillingtagen til, hvor og hvordan algoritmer må benyttes en nødvendighed. Altså må et regelsæt opstilles, der kan begrænse muligheden for misbrug, fejlvurderinger samt risiko for diskrimination. Dog er det ikke kun ved brugen af algoritmer i en retslig sammenhæng, hvor der kan opstå problematikker såsom diskrimination, manglende transparens eller andre biased beslutningstagerer - dette forekommer ligeledes ved

domfældelse på baggrund af den menneskelige fornuft alene. Her bliver spørgsmålet altså, hvorvidt det ene kan betragtes bedre end det andet, eller om en kombination af de to vil skabe de bedste betingelser for en retfærdig dom.

Menneskeliv er en uvurderlig størrelse og derfor er det nødvendigt at der tages vare for dem. Hvis et retssystem skal have tilføjelser; skal systemet forbedres, gøres mere retfærdigt og ikke omvendt.

2.0 Problemformulering

Er det etisk acceptabelt at benytte intransparente algoritmer til risikovurdering i retssystemet?

3.0 Forskningsoversigt

Vi har udvalgt en række argumenter, som bearbejder de etiske problematikker, der medfølger, ved valget af benyttelse af algoritmer til domsafgørelse, frem for domme afgjort udelukkende på menneskelig fornuft. Nedenfor ønsker vi at liste en kort række artikler og tekster, som vil blive taget i brug i vores rapport. Her vil der kort blive redegjort for deres indhold, baggrund mm., samt relevans for vores rapport.

3.1 Julia Angwin, Surya Mattu, Lauren Kirchner Et al.: *Machine Bias*

ProPublica; er artiklen, der først dannede interessen for vores valg af emne til denne rapport. Artiklen er skrevet, udarbejdet og undersøgt af Julia Angwin, som er en prisvindende efterforskningsjournalist og har en B.A. i matematik ved University of Chicago og en MBA ved Graduate School of Business of Columbia University (ProPublica, a, 2019).

Angwin samarbejder med, Surya Mattu; ingeniør og journalist, som har en kandidat ved New York University' Interactive Telecommunications Program og en universitetsgrad ved Nottingham University og journalist (ProPublica, b, 2019). Lauren Kirchner, har ligeledes været med i udvikling af denne artikel. Kirchner har en B.A. i filosofi fra Wesleyan University og en M.S. i journalistik

fra Columbia University. Denne kombination af uddannelser og viden, skaber derfor en stærk artikel, der også blev nomineret til en Pulitzer pris (ProPublica, c, 2020). Artiklen har dannet grundlag for data og statistik til dette problem og oplyser gennem individuelle historier fra det virkelige retssystem, hvilket kan belyse problemet ved COMPAS. Artiklen er altså med til at skabe fundamentet for rapporten, og med denne har vi tilegnet os grund information om, hvor stort den reelle problematik kan anses for at være.

3.2 Alyssa M. Carlson: The Need for Transparency in the Age of Predictive Sentencing Algorithms

Carlson fik først en bachelor i engelsk og musik ved University of Iowa, for derefter at få en interesse for jura og gennemføre en J.D. ved University of Iowa College of Law, hvor hun her ligeledes skrev *The Need for Transparency in the Age of Predictive Sentencing Algorithms* (Octomaslaw, 2019). I denne diskuteres og argumenteres der for, at private firmaer skal have samme standard for transparens som staten også skal, og derved ikke benytte sig af argumentet, om at det er en forretningshemmelighed. Carlson beskæftiger sig til dels med de samme emner som Chiao, dog skelner Carlson sig ved at kigge på et privat firma vs. stat samt dens transparens.

3.3 Vincent Chiao: Transparency: Are Judges Better Than Algorithms?

Vincent Chiao, Ph.d. i filosofi ved Northwestern University og J.D i jura ved Harvard University, arbejder her inden for strafferet, hvilket giver ham stor troværdighed (University of Toronto, ukendt), for teksten vi arbejder med. Chiao undersøger, blandt andet, det etiske spørgsmål om, hvorvidt algoritmerne er intransparente sammenlignet med mennesker, og hvorvidt den beslutning, der bliver taget, er mere transparent og korrekt i forhold til den anden mulighed og dennes beslutning.

3.4 Danielle Kehl, Priscilla Guo, Samuel Kessler: Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing

Denne artikel er skrevet af tre forfattere, den første er Danielle Kehl, som er B.A. fra Yale og studerer jura ved Harvard Law School (Internet Law & Policy Foundry, ukendt). Dernæst er der Priscilla Guo, der har en B.A. i Teknologi, politik og samfund, og en B.A. i kvinder, køn og seksualitets studier ved Harvard Universitet. Den sidste forfatter til artiklen er Samuel Kessler, der er B.A. i historie ved New York University og M.A. og Ph.d. i religiøse studier ved University of North Carolina at Chapel Hill (Kessler, ukendt). Denne brede vifte af folk og deres uddannelser samt kompetencer, er med til at skabe en tværfaglig artikel, der blandt andet giver indsigt i casen; Loomis sagen, der tages i brug længere nede i rapporten. Grundet dens tværfaglighed og fokuspunkter, finder vi den derfor yderst informativ og relevant.

3.5 Cynthia Rudin: Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead

Rudin; professor i Datalogi, elektrisk og computer ingeniør og statistisk videnskab ved Duke University. Rudin's hovedfokus er 'machine learning' og har derfor udviklet '*Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*' (University of Duke, 2020). Denne undersøgelse redegør for, at algoritmen er en 'black box model', hvilket betyder at man forstår mekanismerne bag COMPAS, men selve koden og den egentlige virkning er ukendt. Rudin er et relevant valg, da hun fremviser COMPAS-algoritmen med en datalogisk baggrund, og her tilbyder en unik forståelse for, hvad der er og ikke er realistisk at opnå ved en algoritme som COMPAS, i forhold til en forståelse af mekanismerne bag algoritmen, samt virkningen ved brugen af algoritmen. Rudin kan derfor tilbyde vores rapport en forståelse for, hvordan algoritmen og mekanismer heri virker.

3.6 Jesper Ryberg: Sentencing and Algorithmic Transparency

Jesper Ryberg er Professor i etik og retsfilosofi ved Roskilde Universitet og formand for 'Research Group of Criminal Justice Ethics of Neuroethics and Criminal Justice' (Ryberg, ukendt), og I forlængelse af dette, også vejleder for denne rapport. Vi har valgt at benytte os af Rybergs egne tekster i denne rapport. Ryberg præsenterer tre forskellige argumenter for; transparens-, juridisk-forklaring- samt opacitetsargumentet. Disse tre argumenter, er argumenter som vi i gruppen ligeledes ønsker at berøre i denne rapport. Dette er med til at skabe et stabilt grundlag, at have denne tekst med, da der er fordele ved at kunne diskutere med forfatteren selv.

3.7 William Seymour: Detecting Bias: Does an Algorithm Have to Be Transparent in Order to Be Fair?

Teksten *Detecting Bias: Does an Algorithm Have to Be Transparent in Order to Be Fair?* Er forfattet af William Seymour. Seymour, DPhil Datalogi studerende ved University of Oxford, og en del af Center for Doctoral Training i cybersikkerhed (Department of Computer Science, ukendt). Seymour har, blandt andet, et stort fokus på interaktionsdesign og cybersikkerhed. Dette skaber også hans relevans til denne rapport, da hans store fokus er transparens og om algoritmen behøver dette, for at være fair og kan derfor være med til at bidrage til diskussionen med en anden vinkel, her også sat op med et mere filosofisk perspektiv.

4.0 Redegørelse

4.1 COMPAS-algoritmen og hvordan denne bruges i det amerikanske retssystem.


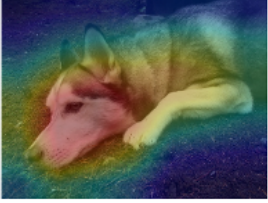

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) er en algoritme, der benyttes hyppigt i det amerikanske retssystem til risikovurdering af kriminelle. Udfaldet

af algoritmens vurdering anvendes af dommere til beslutningen af eventuel prøveløsladelse eller kaution (Rudin, 2019). COMPAS-algoritmen er en såkaldt black box model.

Ved en black box model forstås at de indre mekanismer, eller de indre beregninger algoritmen foretager sig, på sin vis er ukendte. Vi ved med andre ord ikke præcis, hvordan eller hvorfor algoritmen kommer frem til den pågældende risikovurdering og derfor er den intransparent (Rudin, 2019). Det foregående er den klassiske komplicerede forståelse af en black box model. Rudin anvender black box begrebet på to måder, den klassiske komplicerede, og hvad hun kalder den proprietære. COMPAS hører, ifølge Rudin, til den proprietære fremfor den komplicerede. COMPAS er ikke en Machine Learning (ML) model idet den ikke er lavet som en standard ML algoritme. I stedet er algoritmen designet af eksperter baseret på omhyggeligt udviklede spørgeskemaundersøgelser og ekspertise indenfor området (Rudin, 2019). COMPAS' mangel på transparens skyldes i højere grad at udviklerne Northpointe ikke ønsker at offentliggøre koden bag COMPAS, hvilket der kan være flere årsager til; bl.a. at Northpointe er en forretning i en branche med konkurrenter, og derfor ikke ønsker at dele forretningshemmeligheder med konkurrenterne.

Et af de største argumenter bag fortsat brug af black box ML modeller er at, en intransparent model ofte er mere kompliceret og dermed mere præcis. Hvorimod en transparent model vil være nødsaget til at være simple, for netop at opretholde sin transparens, og dette går udover præcisionen. Altså er der et formodet kompromis mellem præcision og transparens. Derfor anvendes intransparente ML modeller stadig i stor stil, men et forsøg på at forklare, hvordan disse fungerer, vil ofte ikke give mening eller ikke være fyldestgørende nok (Rudin, 2019).

Rudin anvender billede bearbejdning, såkaldte *saliency maps*, til at demonstrere, hvorfor det er vanskeligt at forstå, hvordan en ML model fungerer i praksis. saliency maps anvendes ofte som demonstration/forklaring for, hvordan modellen når sit resultat, men som Rudin demonstrerer, yder saliency maps ingen egentlig forklaring bag modellens beregning. Et saliency map giver blot information om, hvilken del af billedet modellen bearbejder, og ikke hvad modellen reelt gør med denne information. Samtidig påstår Rudin at der er en uheldig trend, til kun at demonstrere eksempler på saliency maps, som giver det korrekte resultat. Forstået på den måde at ML modeller selvfølgelig ikke er perfekte, men laver fejl, og det samme billede kan blive fortolket af modellen som to forskellige ting som det kan ses i figur 1 (Rudin, 2019).

	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps			

Figur 1 (Rudin, 2019:5)

Af figur 1 fremgår det, hvordan selvsamme billede bliver fortolket af en ML model til at være to forskellige ting. Samtidig kan det ses, hvilke dele af billedet modellen bearbejder. Det er dog ikke klart ud fra dette saliency map, hvordan modellen analyserer det pågældende data, således at den i det ene tilfælde kommer frem til det rigtige resultat, men samtidig, i det andet tilfælde, kommer frem til et forkert resultat. Hvis blot det korrekte resultat bliver præsenteret, er der fare for, at dette kan give en falsk fornemmelse af, hvordan en ML model fungerer. Selvom detaljerne ikke står klart, betyder det måske ikke nær så meget når resultatet er rigtigt. Det er dog tydeligt at der ikke er tale om en reel forståelse af, hvordan modellen fungerer (Rudin, 2019).

Dette er et eksempel som Rudin anvender til at demonstrere komplikationen i at forklare og forstå ML modeller. Pointen er, at black box modeller ikke kan forklares fyldestgørende og i detaljer. Derfor er det ikke er klart, hvorfor modellen i nogle tilfælde kommer frem til et korrekt resultat og i andre tilfælde kommer frem til et forkert.

COMPAS-algoritmen er dog, ifølge Rudin, ikke intransparent grundet dens kompleksitet, men snarere af proprietære årsager. Dette er altså som nævnt tidligere, den primære årsag til COMPAS-algorithmens intransparens, og hvorfor det ikke er offentlig kendt hvordan algoritmen i praksis fungerer (Rudin, 2019).

Vi ved dog at COMPAS-algoritmen risikovurderer den tiltalte ved hjælp af et pointsystem mellem 1-10. Den enkelte tiltalte tildeles point baseret på mere end 100 faktorer, heriblandt køn, alder og tidligere begået kriminalitet. Race er ikke en af de anvendte faktorer (Corbett-Davies, 2016). Indsamlingen af data til COMPAS-algoritmen, for den enkelte tiltalte, foregår som en form for spørgeskemaundersøgelse, hvor den tiltalte selv svarer på nogle spørgsmål og andre svar hentes fra straffeattesten. Spørgeskemaet indeholder spørgsmål omkring: familiemedlemmers eventuelle fængselstid og hvorvidt den tiltalte havde problemer med andre børn i sin skoletid. Samtidig bliver

den tiltalte bedt om at tage stilling til forskellige udsagn, eksempelvis om en person, som sulter, har ret til at stjæle (Angwin, Mattu, Kirchner Et al., 2020). For et komplet eksempel på et COMPAS-spørgeskema se bilag 1.

COMPAS-algoritmen gør også brug af mere dynamiske faktorer; hvilke og hvordan disse bruges, er mere uklart, af Northpointe selv fremgår det blot:

”COMPAS relies on both static and dynamic data to generate its risk and needs results. The use of dynamic measures allows for measures to change over time as behavior changes. These changes are included in the measures of risk and need. The dynamic factors also allows for the “overlay” of previous assessments on the latest assessment to visual see any change in risk and need scores.” (Northpointe inc., 2012: 1).

Det er altså kun muligt for os at redegøre for COMPAS-algorithmens forskellige komponenter og ikke for, hvordan disse spiller sammen, eller for hvilke faktorer, som vægtes højere end andre, og vigtigst af alt, hvordan algoritmen i sidste ende kommer frem til den pågældende risikovurdering. Dette gør sig samtidig også gældende for de dommere, rundt omkring i det amerikanske retssystem, der får præsenteret disse risikovurderinger som et værktøj til deres beslutningstagen omkring eventuel prøveløsladelse. Ikke desto mindre anvendes COMPAS og andre lignende værktøjer bredt i det amerikanske retssystem. Der er selvfølgelig forskel på, hvor stor indflydelse algoritmens risikovurdering har fra stat til stat og på den individuelle dommers beslutning. Ikke desto mindre er der tilfælde, hvor algoritmens vurdering har stor indflydelse, ikke blot på den tiltales mulighed for prøveløsladelse, men også på længden af selve fængselsstraffen. Særligt i Wisconsin bliver COMPAS-algoritmen anvendt på denne måde:

”In theory, judges are not supposed to give longer sentences to defendants with higher risk scores. Rather, they are supposed to use the tests primarily to determine which defendants are eligible for probation or treatment programs.

But judges have cited scores in their sentencing decisions. In august 2013, Judge Scot Horne in La Crosse County, Wisconsin, declared that defendant Eric Loomis had been “identified, through the COMPAS assessment, as an individual who is at high risk to the

community.” The judge then imposed a sentence of eight years and six months in prison.”
(Angwin, Mattu, Kirchner et al., 2020)

Af dette eksempel er det tydeligt at se COMPAS-algorithmens mulige indflydelse på omfanget den enkelte tiltaltes straf. Samtidig er algorithmens indre mekanismer intransparente, og det er derfor tydeligt, hvorfor der opstår en problematik her, når undersøgelser peger i retning af at COMPAS-algoritmen er urimeligt partisk. I resten af rapporten ser vi nærmere på og diskuterer, problematikken bag en så indflydelsesrig, men intransparent algoritme.

4.2 Rudins argumentation mod brugen af black box ML modeller

Som nævnt i redegørelsen for COMPAS-algoritmen skelner Rudin mellem to former for black box modeller; den klassiske komplicerede form for black box modeller, og den proprietære (se afsnit ovenfor). Den klassiske black box model er intransparent på grund af dens kompleksitet. Disse ML modeller er så avancerede og komplicerede, at vi som mennesker simpelthen ikke forstår, hvordan de fungerer, eller med andre ord; hvordan/hvorfor algoritmen kommer frem til det pågældende resultat. Den proprietære black box model er intransparent da virksomheden, som har udviklet algoritmen, ikke ønsker at løfte sløret for algoritmens beregninger bag dens resultat (Rudin, 2019).

Rudins formål med artiklen er at flytte overbevisningen fra den nuværende forståelse af, at en black box er en nødvendighed for præcise forudsigelser, og over til det, der er Rudins hovedargument; at black box modeller ikke skal accepteres trods deres intransparens, med den fejlfortolkning at de opvejer for dette ved argumentet om at være mere præcise. Rudin argumenterer for at mere transparente modeller kan være lige så præcise, og at der derfor er nødt til at ydes en markant indsats i at udvikle sådanne modeller, frem for fortsat at benytte intransparente black box modeller (Rudin, 2019).

En fordel ved simple, men mere transparente modeller er, at det er muligt at analysere resultaterne af disse i modsætning til en mere kompliceret black box model. Rudin argumenterer ligeledes for, at ved undersøgelse i felten, hvor forudsigelses algoritmer anvendes, vil den simple mere transparente model i sidste ende komme frem til et mere præcist resultat. For som Rudin

påpeger, kræver det som regel en iterativ proces for at skabe viden fra indsamlet data. Denne proces er naturligvis langt nemmere at foretage på den simple transparente model, frem for den komplicerede intransparente model (Rudin, 2019).

”Generally, in the practice of data science, the small difference in performance between machine learning algorithms can be overwhelmed by the ability to interpret results and process the data better at the next iteration [19]. In those cases, the accuracy/interpretability tradeoff is reversed – more interpretability leads to better overall accuracy, not worse.” (Rudin, 2019: 2f.).

Selv når der fremstilles forklaringsmodeller, som et forsøg på at forklare black box modeller, så fejler disse, og hvis de fejler, skaber det blot mistro til den algoritme som modellen skulle forsøge at forklare. Hvis en model, til at forklare en black box model, har ret i 90% af tilfældene, så tager den stadig fejl i 10% af tilfældene og derfor kan der ikke stoles på den eller den originale black box. Desuden kan det anses problematisk at anvende et begreb som at modellen skal ‘forklare’ en black box model. Ifølge Rudin er der nærmere tale om en approksimation (Rudin, 2019).

I henhold til COMPAS-algoritmen og ProPublicas analyse af denne, påpeger Rudin en problematik i brugen af ovenstående begreb. Den model som ProPublica har opstillet for at forklare COMPAS-algoritmen er muligvis præcis i at efterligne algoritmens resultater, men den er ikke trofast til COMPAS-algorithmens originale beregning, eftersom ProPublica anvender race direkte i deres model og COMPAS ikke gør (Rudin, 2019):

”Most recidivism prediction models depend explicitly on age and criminal history, but do not explicitly depend on race. Since criminal history and age are correlated with race in all of our datasets, a fairly accurate explanation model could construct a rule such as “This person is predicted to be arrested because they are black.””(Rudin, 2019: 4).

Dette skaber en tydelig problematik både i ulemperne ved en intransparent black box model, men også ved at fremstille såkaldte forklaringsmodeller, som i virkeligheden måske ikke giver den originale model en retfærdig forklaring. Rudin forsøger altså med dette citat at slå fast, hvorfor det er nødvendigt med et oprigtigt forsøg på at skabe nye transparente modeller. For at problematikker

som disse førnævnte, ikke behøver at opstå eller er i hvert fald er til at fortolke og analysere nærmere på, måske i sidste ende komme frem til en endnu mere præcis vurdering. Hvis ikke der sker et skift mod fremstilling af mere transparente modeller, er det muligt at brugen af black box modeller blot vil fortsætte, selvom det ikke nødvendigvis anses for at være forsvarligt at benytte dem. Det er især nødvendigt med et skift mod mere transparente modeller, når det ikke er klart, hvad der kan godtages som forklaring for en vurdering foretaget af en black box (Rudin, 2019).

4.3 COMPAS' intransparens

I dette afsnit, vil der blive fremlagt plausible grunde til, hvorfor COMPAS-algoritmen er intransparent og, hvorfor det ikke umiddelbart er ligetil at gøre algoritmen transparent, som der er blevet fremlagt et ønske om, fra diverse instanser. Disse instanser nævnes i rapporten.

Rudin argumenterer for, at grunden til at COMPAS ikke er transparent, er fordi det er en forretningshemmelighed, og den derfor er lavet til at profitere (Rudin, 2019). Som så meget andet, ville det ikke være en god forretning, hvis alle har adgang til, hvordan den virker og hvordan den er lavet, da konkurrenter af denne årsag ville kunne efterligne den og eventuelt tjene penge på den. Rudin argumenterer ligeledes for at en transparens kan være med til, at motivere firmaerne og lade dem konkurrere internt i forsøget på producere et bedre produkt, hermed vil der ikke blive skabt monopol, "Thus, transparency could help improve the quality of the system" (Rudin, 2019: 7).

Selvom det kan virke som et aktivt valg, fra Northpointes side, at skabe monopol på COMPAS-algoritmen og hermed bibeholde sine kunder, er en anden vigtig faktor; kompleksiteten af selve algoritmen. For at kunne lave en algoritme, så er det en nødvendighed at være i stand til at programmere, det kræver derfor mange års erfaring. Selvom en algoritme gøres transparent, vil der stadig være den udfordring, at skulle have indsigt i programmering og datalogi generelt. For at kunne forholde sig kritisk til algoritmens risikovurdering, er det derfor nødvendigt for dommere, jurister og anklagere at opnå en fundamental forståelse for algoritmens virken.

Kan et menneske opnå fuld forståelse for en algoritme? Der er mange ting under datalogien, et menneske dårligt forstår grundet dens kompleksitet. Specielt ved en selvlerende algoritme der konstant er i udvikling, som William Seymour beskriver i sin tekst, så er dennes kompleksitet også dets styrke; "many machine learning methods are useful precisely because they work in a way

which is alien to conscious human reasoning.” (Seymour, 2018: 1). Seymour opstiller derfor betingelserne for, at der enten kan benyttes sig en nem, dog mindre effektiv algoritme som er til at forstå, eller bruge en mere kompleks, dog en algoritme, der er sværere at forstå. En stræben efter transparens skal ikke opgives, selvom dette kan være skyld i, at produktet ikke opnår det optimale. Derimod skal en mere præcis vurdering af, hvad der efterspørges og er brug for, identificeres (Seymour, 2018).

4.4 Carlsons argumentation for transparens i risikovurderings algoritmer

Carlson ønsker at argumentere for; nødvendigheden af transparens af algoritmer som anvendes i retssystemet. Ifølge Carlson er det vigtigt, at stater i inkorporationen af nye teknologier, heriblandt risikovurderings algoritmer, ikke går på kompromis med værdierne omkring en åben og transparent stat. Det anses derfor problematisk når retssystemet, som repræsenterer staten, tager proprietære værktøjer som COMPAS i brug. Samtidig argumenterer Carlson for at disse værktøjer ikke undersøges dybdegående nok, før de tages i brug. Staterne har altså ikke tilstrækkeligt belæg, for modellernes præcision, før de begynder at drage konklusioner ud fra disse modeller. Ligeledes betyder dette, at tiltalte såvel som forsvarer, er nødsaget til at acceptere den pågældende dom, uden mulighed for at teste præcisionen eller validiteten af algoritmen (Carlson, 2017).

En spørgeskemaundersøgelse fra 2010 fandt frem til, at næsten alle stater anvender en form for risikovurderingsværktøj. Dette gøres naturligvis i et forsøg på at bedre håndtere mange sager, desuden har risikovurderingsværktøjer været anvendt helt tilbage fra 1927. Det er derfor et helt naturligt element af en retssag i dag, og risikovurderings algoritmer er en branche med megen konkurrence blandt udviklere. Dette har dog den meget uheldige konsekvens at virksomheder som Northpointe holder beregningerne/vægtningen af faktorer i deres værktøj hemmeligt, for bedst muligt at kunne udkonkurrere konkurrenterne og ende med profit (Carlson, 2017).

Dette er et problem når mange stater samtidig ikke udfører et tilfredsstillende arbejde i et forsøg på at validere præcisionen af de algoritmer som de anvender. Især når det ses mere og mere tydeligt at algoritmens vurdering har direkte indflydelse på dommerens vurdering/beslutning: “In an appeals hearing, Judge Babler explained his sentencing decision: “Had I not had the COMPAS,

I believe it would likely be that I would have given one year, six months.”” (Carlson, 2017: 319f.). Dette er blot et eksempel på en sag, hvor den tiltalte endte med en hårdere dom, nemlig to år, frem for halvandet år som var det dommeren oprindeligt havde tænkt sig at give, men pga. COMPAS-algorithmens vurdering lod dommeren dette påvirke sin egen vurdering. Hertil bliver det jo ekstra problematisk at den dømte ikke har mulighed for at få svar på, hvad der gør at algoritmen vurderer ham som højere risiko end dommeren oprindeligt ville have gjort, han er blot nødsaget til at acceptere sin dom (Carlson, 2017). For et andet mere dybdegående eksempel se afsnittet omkring The Loomis case.

Den helt store problematik for Carlson er at flere steder bliver værktøjer som COMPAS implementeret uden først at teste validiteten, samtidig med at disse værktøjer er intransparente. Ifølge Carlson så stemmer dette ikke overens med de værdier omkring en transparent og ansvarlig stat, som er en nødvendighed i et fungerende demokrati. Firmaerne som supplerer staten med disse værktøjer, får lov til at holde deres modeller proprietære grundet kommerciel lovgivning. Carlson mener at når virksomheder som disse vælger at udvikle værktøjer til brug af den offentlige stat, så bør disse værktøjer imødekomme samme forventninger om transparens og ansvar, som ellers forventes af staten. Ideen om at lade virksomheder holde deres konkurrenceevne intakt ved ikke at afsløre de indre mekanismer af deres produkt, kan ikke gøre sig gældende for produkter som COMPAS, for virksomhedens interesse kan ikke veje tungere end offentlighedens interesse (Carlson, 2017).

4.5 Om straffeprocess

Den følgende beskrivelse af en straffeprocess vil ikke gå i tekniske detaljer med en straffeprocess' delelementer, men i stedet fokusere på et særligt punkt. Det drejer sig om, hvorvidt en dommer må kende til oplysninger om tiltalte som juridisk kan tilbageholdes fra tiltalte. Dommeren må altså træffe en domsafsigelse på et grundlag, der kan være kendt af dommeren, men ikke af tiltalte.

Når den tiltalte skal modtage sin dom, vil dommeren, eller den givne autoritet, anmode om en præ-doms undersøgelsesrapport (PSI-rapport). Denne vil indeholde information om den tiltaltes liv og baggrund. Præ-doms rapporten bliver samlet af en juridisk medarbejder med baggrund i socialt arbejde. Rapporten kan indeholde information om den tiltaltes straffeattest, detaljer fra interviews med familie, venner og tidligere arbejdsgivere samt personlige og biografiske detaljer. Fra lovens

side er det utroligt begrænset, hvad denne rapport ikke må indeholde (Kehl, Guo, Kessler, 2017). Der opstillet et strengt sæt regler for, hvad der må introduceres af bevismateriale under en retssag, men her er det faktisk tilladt for dommeren at forholde sig til langt mere materiale end blot, hvad der introduceres under selve retssagen, og sågar hvad den tiltalte er tilladt adgang til. Rationalet bag denne beslutning er bygget på ideen om, at en tiltalt ikke skal dømmes udelukkende på den givne kriminelle handling. Den tiltaltes liv og karakteristika har ligeledes indflydelse (Kehl, Guo, Kessler, 2017). Altså skal sagerne behandles individuelt, da dette medfører en mere fair retssag.

Når PSI rapporten er færdiggjort, overleveres den til dommerens skue - selvom denne information ligeledes overleveres til den tiltalte, kan det forekomme, at en vis mængde information vil blive tilbageholdt og den tiltalte vil ikke nødvendigvis kunne få et fuldt overblik over de informationer, dommeren har til rådighed. Dette er retfærdiggjort ved, at hvis information er for følsom, for eksempel; hvis det er vigtig information fra en pårørende til den tiltalte, så kan informationen være afgørende og derfor skal sagen kunne tilbageholdes, da den pårørende ikke føler sig tryk ved sådan en type udtalelse. Derfor tilbageholdes dette fra den tiltalte (Kehl, Guo, Kessler, 2017). Dommeren er fri til at gøre med denne rapports information som han/hun finder nødvendigt for en fair dom, også selv om informationen ikke er fremlagt under retssagen.

4.6 The Loomis case: Ethiske komplikationer ved COMPAS-algoritmen

The Supreme Court of Wisconsin, 2016. Her findes den første store sag mod brugen af risikovurderings algoritmer til domfældelse. Denne retssag adresserer nogle af de bekymringer, der er opstået i forhold til dommeres benyttelse af denne type værktøj, til blandt andet at vurdere længden på en straf - altså en mere direkte indflydelse på straffen selv. Her bliver problematikken, om dette er en overtrædelse af de rettigheder som samfundet har, samt om der med denne algoritme medfølger en diskrimination af særlige individer (Kehl, Guo, Kessler, 2017).

Eric Loomis, den tiltalte i retssagen, blev i sin tid anholdt for at sidde bag rattet i et drive-by-shooting. Han ender med at erklære sig skyldig i handlingen - retten efterspørger herefter en PSI-rapport, hvori en risikovurdering lavet af COMPAS-algoritmen, er vedlagt. Loomis er vurderet til at være i high-risk kategorien - her både i sandsynligheden for tilbagefald til generel kriminalitet,

men ligeledes i high-risk for voldelige handlinger (Kehl, Guo, Kessler, 2017: 18). Loomis er tidligere dømt som sexforbryder, hvilket menes at have haft indflydelse på algoritmens vurdering af ham som high-risk.

På baggrund af denne vurdering, samt andre beviser, modtager Loomis en dom på 8 år og 6 måneder, med udtalelsen: “The risk assessment tools that have been utilized suggest the you’re extremely high risk to reoffend” (Kehl, Guo, Kessler, 2017: 18).

Efter at have modtaget sin dom protesterer Loomis mod dommen, han begrundet dette på algoritmens indflydelse, som en strid mod hans basale rettigheder; altså at han ikke har modtaget en fair retssag, da han har retten som individ til at blive dømt på præcise fakta og beviser - dette mener Loomis med COMPAS involveret i sagen, ikke er muligt. Algoritmen er ikke åben og tilgængelig for inspektion og derfor kan man ikke være sikker på årsagen til det givne resultat (Kehl, Guo, Kessler, 2017).

Endvidere udfordre Loomis sin dom; Han mener ikke, at hans ret til en individuel retssag er bliver overholdt, retssagen skal afholdes og vurderes ud fra det enkelte individ - hvilket Loomis argumenterede for at COMPAS-algoritmen ikke var i stand til, da den er bygget op omkring og tager udgangspunkt i information baseret på generelle befolkningsmæssige karakteristika – altså på en større gruppering af mennesker - til at vurdere, hvorvidt hans tilbøjelighed er for fremtidig kriminalitet (Kehl, Guo, Kessler, 2017). Ligeledes anklager Loomis COMPAS-algoritmen for at være diskriminerende, da algoritmen tager højde for ‘køn’ i sine beregninger.

Flere argumenter i denne retssag ender med at være fremtrædende, herunder Loomis’ argument om ‘nøjagtighed’ - her italesættes en problematik vedrørende algoritmens arbejde, arbejder den præcist nok? Loomis’ mener at man ikke kan benytte algoritmen, da man ikke har adgang til dennes data og dermed ikke kan afgøre om resultatet er tilstrækkeligt og nøjagtigt nok. Retten ender med at afvise denne påstand, de medgiver at der kan være en manglende tilgængelighed af COMPAS’ data, men at dette ikke har været den afgørende faktor for hans dom - Algoritmen har benyttet sig af et spørgeskema, som Loomis’ selv har udfyldt og af denne grund, har han, som tiltalt, haft tilgængelighed nok til dataen og derfor er dommen på baggrund af dette valid (Kehl, Guo, Kessler, 2017).

I henhold til Loomis argument om; et brud på hans ret til en retssag baseret på individualitet, bliver afvist på begrundelsen af algoritmens betydning i forhold til dommerens brug. Argumenterer retten for, at algoritmens indflydelse på selve dommen ikke har været den afgørende faktor for resultatet

hvis dette ikke havde været tilfældet, altså at COMPAS havde været den afgørende faktor for Loomis' dom, så havde retten måske accepteret hans argument (Kehl, Guo, Kessler, 2017).

Retten ender ligeledes med at afvise Loomis' argument om algoritmen som diskriminerende, her på baggrund af en kønsforståelse. Dette gør de ved at argumentere for, at køn i denne sammenhæng faktisk er en relevant faktor og en vigtig faktor at tage højde for, da der er forskel på mænd og kvinders sandsynlighed for tilbagefald, samt mulighed for rehabilitering og derfor er algoritmens køns-vurdering med til at øge muligheden for en retfærdig retssag (Kehl, Guo, Kessler, 2017: 19).

Retten ender med at afvise alle Loomis' protester, og fastholder at brugen af COMPAS-algoritmen har været forsvarlig at benytte i forbindelse med domsfældning - dog eftergiver retten, at brugen af algoritmen fremover skal have større restriktioner; algoritmen har behov for restriktioner, samt at denne type værktøj ikke er egnet til at skulle benyttes i direkte forbindelse med varigheden eller hårdheden af en given straf. Her på grundlag af, at COMPAS ikke er udviklet specifikt til dette formål. Fremover skal den anses for at være egnet til vurdering af hvorvidt Low-risk kriminelle skal forblive fængslet eller om der er mulighed for prøvelødsledelse og hvad dette yderligere skal indebære:

- “1. COMPAS is a proprietary tool, which has prevented the disclosure of specific information about the weights of the factors or how risk scores are calculated;
2. COMPAS scores are based on group data, and therefore identify groups with characteristics that make them high-risk offenders, not particular high-risk individuals;
3. Several studies have suggested the COMPAS algorithm may be biased in how it classifies minority offenders;
4. COMPAS compares defendants to a national sample, but has not completed a cross-validation study for a Wisconsin population, and tools like this must be constantly monitored and updated for accuracy as populations change; and
5. COMPAS was not originally developed for use at sentencing.” (Kehl, Guo, Kessler, 2017: 18f.)

Kehl, Guo og Kessler fremfører en interessant pointe i denne forbindelse, hvordan skal en dommer bruge dataen fra algoritmen, hvis ikke at denne må have indflydelse på dommen?

5.0 Analyse & diskussion af centrale argumenter

Vi vil nu se på argumenter for, hvorfor man ikke bør bruge intransparente algoritmer. Argumenterne har hver især et specifikt fokuspunkt i den juridiske proces, eller i algoritmers tekniske opbygning. Vi vil i det følgende gennemgå argumenterne og diskutere om deres præmisser grundlæggende er sande og dermed om argumentet kan føres. Derudover vil vi diskutere indvendinger mod argumenterne og se på, i hvor høj grad, argumenterne er plausible.

5.1.0 Første argument: Intransparente algoritmer bør ikke anvendes, da det vil skabe mistillid i befolkningen

Et argument for, hvorfor transparens er vigtigt, er at man ikke bør bruge intransparente risikovurderings algoritmer fordi det vil skabe øget mistillid i befolkningen, i forhold til retssystemets måde at fælde domme på (Ryberg, 2020: 14).

Manglende transparens skaber mistillid, hvis befolkningen eksempelvis ikke har forståelse for, hvordan algoritmen bidrager til dommerens vurdering af dom, vægtning af inputs eller adgang til algoritmens kode.

Det er afgørende for samarbejdet mellem befolkning og retssystem at der er tillid mellem de to parter, hvis ikke det er opnået vil det kunne underminere hele formålet med systemet. Eksempelvis kan mistillid til retssystemet gøre at befolkningen ikke vil stille op som vidne i en retssag, eller melde general kriminalitet, da befolkningen ikke vil se noget formål med dette.

5.1.1 Diskussion af argumentet

Vi vil nu se på tre indvendinger mod argumentet om, at intransparens skaber mistillid i befolkningen. Er det virkelig sådan, at hvis man bruger intransparente algoritmer, så skaber det mistillid? For at svare på denne indvending, så er det nødvendigt at få afklaret nogle forhold først. Først må det afklares, hvilken transparens vi taler om, er det transparens af koden?

5.1.2 Første indvending: Transparens af koden er irrelevant, da befolkningen alligevel ikke forstår den

Man kan indvende at befolkningen slet ikke ville sætte sig ned og læse koden, da den ikke er let at forstå. Så det er svært at forestille sig, at det ville skabe mistillid, at algoritmerne er intransparente i forhold til deres kode;

“when reviewing even a moderately-sized code base, it quickly becomes apparent that issues of transparency and interpretability cannot be resolved simply by making computer code available” (Seymour, 2018: 5 i pdf).

Umiddelbart så tyder det på, at algoritmens opbygning og kodning er så kompleks at selv, hvis befolkningen fik adgang til koden, så ville det være så svært at forstå, hvordan den fungerer, at det ikke ville skabe mere eller mindre tillid til retssystemet. Det kan altså ikke ligefrem siges at være sandt, at intransparens skaber mistillid, hvis vi blot tænker på intransparens i forhold til algoritmens kode.

Denne indvending fører os videre til en diskussion af et argument for at intransparens skaber mistillid, hvis algoritmens kode er beskyttet og skjult som en forretningshemmelighed. Det er netop, hvad der gør sig gældende for COMPAS-algoritmens intransparens, da den er beskyttet under lovgivningen om at det er en forretningshemmelighed. Det er lovligt ikke at offentliggøre en algoritmes kode med argumentationen om at det er en forretningshemmelighed, netop fordi det er måden koden er skrevet på, der gør at firmaet kan tjene penge på dens evne til risikovurdering,

så kan det forholdsvis let forhindres. Det gøres ved at fjerne denne form for lovgivning og sørge for open access til algoritmens programmering (Ryberg, 2020).

Spørgsmålet bliver her om intransparens på dette grundlag, vil skabe mistillid i befolkningen og derfor ikke bør anvendes? Hvis befolkningen har en forventning om at staten optræder transparent, men staten så alligevel anvender risikovurderingsværktøjer, der er intransparente, grundet forretningshemmeligheder, fremstår staten nu som intransparent og derfor kan brugen af proprietære algoritmer skabe mistillid blandt befolkningen. Derfor kan det siges, at hovedargumentet stadig kan betragtes som validt. Er dette tilstrækkeligt nok til at sige, at man ikke må benytte algoritmer, der er intransparente? Hvis man antager at en algoritme er en forretningshemmelighed og at denne sælges til brug i statens retsinstant, så behøver virksomheden i princippet ikke at optimere algoritmen når først den er solgt. Hvis den udgave af algoritmen, som virksomheden har solgt, er diskriminerende, så er svaret; ja, det er tilstrækkeligt at sige, at hvis en algoritme er intransparent fordi den er en forretningshemmelighed, så bør den ikke anvendes.

5.1.3 Anden indvending: Dommere kan også betragtes som intransparente, så hvis intransparente algoritmer skaber mistillid, så må det samme kunne siges om dommere

Vi må starte med en antagelse om, at den generelle befolkning har tillid til deres retssystem. Med denne antagelse, kan der altså argumenteres for, at befolkningen udviser en tillid til dommerne i retssystemet. Dommere kan betragtes som intransparente, da der ikke er nogen forventning til en begrundelse for, hvad der fører til en given beslutning. De kan handle ud fra en vis mængde intuition, samt være påvirket af underliggende forforståelser, som de ikke nødvendigvis selv er klar over (se afsnit om straffeprocess). Ved intransparente algoritmer har vi stadig mulighed for en bedre forståelse for dennes virken, da algoritmen handler ud fra de inputs, den modtager, hvilket programmøren i stort omfang er herre over. Input data kan der netop holdes godt øje med, da den manuelt skal tilføjes til algoritmen gennem eksempelvis spørgeskemaer (jævnfør afsnit om brugen af COMPAS-algoritmen i det amerikanske retssystem). Af denne grund kan det siges, at vi derfor kan forstå algoritmens virken bedre end dommerens egen. Endvidere skal det huskes, at algoritmer

som COMPAS, skal benyttes i samarbejde med dommeren, den skal være et hjælpemiddel og ikke den afgørende faktor - der kan derfor argumenteres for, at selv en intransparent algoritme må kunne være med til at skabe mere tillid i befolkningen, hermed må det kunne betragtes forsvarligt at benytte intransparente algoritmer.

Hvis intransparensen omhandler, hvordan dommere vurderer og forklarer brugen af algoritmer, så kan dette resultere i en mistillid. Der findes en anden type indvending for, hvorfor transparens skaber mistillid i befolkning. Hvis eksempelvis en menneskelig dommer kan forklare, hvordan personen er kommet frem til afgørelsen, så kaldes dette en intelligibel transparens (Chiao, 2020).

Denne intelligible transparens må kunne kritiseres i forbindelse med, at en dommer anvender algoritmiske risikovurderinger. Da algoritmer, der er programmeret til at finde korrelation mellem den input data de får og de forudsigelser algoritmer laver, kan optimere forudsigelsesprocessen selv, for at finde endnu flere korrelationer mellem data, og dermed følge så avanceret en proces at selv ikke programmørerne kan gennemskue, hvordan det fungerer. Så det er svært at antage at dommeren har mulighed for at forklare hele algoritmens proces, men muligvis godt kan forklare dommerens eget valg om at anvende algoritmens forudsigelser, og hvordan denne forudsigelse er blevet anvendt. Det kan altså tyde på at den intelligible transparens, der bliver argumenteret for som eksisterende i forhold til menneskehjernens beslutningsproces, er intransparent selvom dommeren er overbevist om at kunne forklare, hvordan algoritmen er blevet brugt. Lad os se på tre argumenter for dette. For det første, så eksisterer der en stor splittelse mellem de grunde en person giver til, hvorfor de vælger som de gør, og de faktorer, der faktisk forklarer deres valg (Chiao, 2020). Og det er også selvom personen er helt overbevist om selv at kende grunden. Denne splittelse mellem det personen tror er grunden, og det faktiske grundlag for valget, kaldes psykologisk opacitet. En naiv person vil være overbevist om selv at kunne finde grundlaget for sit valg, og modsat ville erfarne personer acceptere denne splittelse og vide at det sande grundlag ikke udelukkende er det, personen selv kan tænke sig frem til (Chiao, 2020).

Dog er det ikke altid at grundlaget for valget skal findes introspektivt i personen selv. Der eksisterer også eksterne faktorer, der spiller ind på, hvordan valget tager sig ud. Eksempelvis er *'framing'* en måde eksternt at påvirke en persons beslutningsproces (Chiao, 2020: 9). Hvis personen bliver stillet overfor det samme valg, men præsenteret på to forskellige måder, så ville personen, hypotetisk, så ville valget i begge scenarier være konsistent (Chiao, 2020).

Et andet argument for den psykologiske opacitet, der eksisterer i en dommers beslutningsproces, er følgende: Der findes en normal ubevidst menneskelig bias, hvor individet bliver forankret ved et bestemt sæt eksplicite informationer. Det kan være at individet føler en ekstra fiksering overfor datapunkter, eller personer, der udviser ekstra sympati, og at dette påvirker, hvordan dommere træffer deres valg (Chiao, 2020). Disse forankringseffekter har især vist sig effektive og indflydelserige når datapunkterne, som forankringen er fastlåst på, er irrelevante. Dette tyder på en stor ubevidst kognitiv bias i beslutningsprocessen:

“Researchers have also found that judges, like laypeople, are also susceptible to anchoring effects, even based on anchors stemming from the interested and/or irrelevant actions of the parties” (Chiao, 2020: 10).

Eksempelvis er tyske dommere blevet præsenteret for en beskrivelse af en hypotetisk forbrydelse som de skulle vurdere. I sagen var der lagt unaturligt meget vægt på visse irrelevante dele af informationerne om tiltalte, for at vægte forankringen på et ikke plausibelt grundlag. Dommerens beslutninger blev derfor i høj grad anset som gæt. Resultatet var at dommerne gav langt strengere straffe, hvis informationerne, som dommerne havde at arbejde med, var irrelevante, og de gav lavere straffe, hvis informationerne var mere relevante i forhold til en vurdering af forbrydelsen (Chiao, 2020). Man kan tænke sig at dommere ikke ved, hvilke faktorer, der påvirker deres beslutning. En af disse faktorer er at algoritmerne lærer sig selv, hvordan de skal vægte og vurdere de input data som de får, så de kan transformeres til output data i form af risikovurderinger (Chiao, 2020). Det vil sige at selv om dommeren ved, hvordan algoritmen fungerer, så er det ikke muligt at bibeholde en forståelse fordi algoritmen hele tiden optimerer og justerer sig selv. Er det nok at sige at, hvis en intransparent algoritme kan optimere og justere sig selv, så må den ikke bruges? Nej ikke nødvendigvis, for hvis algoritmen kan optimere og justere input data, og den bliver endnu mere præcis, så er det en fordel at lade den foretage disse ændringer da dette vil skabe mere præcise risikovurderinger. Hvad så med kritikken af den psykologiske opacitet i dommeres beslutningsproces, kan det siges at dommere ikke må anvendes, hvis de ikke fuldt ud kan redegøre for deres beslutninger? På den ene side kan der argumenteres for at algoritmer kan være mere objektive end dommere, da de ikke har de menneskelige bias som vi lige har set ovenfor. Desuden kan de programmeres til at give så optimale risikovurderinger som muligt. Ville det ikke skabe mere mistillid

i befolkningen, hvis vi helt fjerner dommere fra retssalen og i stedet lader algoritmers risikovurderinger være det fulde grundlag for en domfældelse? Nej, det ville være overdrevet at argumentere for at dommere ikke må benyttes i retssalen, da dommere er med til at skabe tillid i befolkningen ved netop at give tiltalte en følelse af, at det er et menneske, der behandler vedkommendes sag.

Argumentet om at transparens øger tilliden hos befolkningen kan virke plausibel, da transparens giver tiltalte større indsigt i algoritmers måde at virke på, end de havde til at starte med. Det kan virke som om det er en mere fair og åben behandling, men der er en sidste faktor vi er nødt til at se på. Helt præcist er faktoren; hvad udgangspunktet er i forhold til forventningerne til algoritmen:

“For users who have low confidence in a system, with little idea how it works, then transparency can help form working system models, thus boosting confidence and trust. In contrast, trust can be undermined if users have a working theory of the system but are exposed to anomalous behaviors such as system errors” (Springer & Whittaker, 2018: 6.2).

Tilliden afhænger af, hvilke forventninger den tiltalte i udgangspunktet har til algoritmer og at transparens derfor forstærker følelsen af henholdsvis høj eller lav tillid (Ryberg, 2020). Hvis tiltalte har stor tillid til algoritmerne, men ikke har indsigt i, hvordan algoritmerne virker, så kan transparens skabe mistillid når tiltalte får indsigt i algoritmens komplekse opbygning eller vægtning af input. Modsat kan en bruger med dårlig tillid til algoritmer få styrket sin tillid gennem transparens, ved at opdage, hvordan algoritmer kan fungere præcist, og hvordan de kan justeres, så de skaber fair outputs til dommerne. Tillid er til dels afhængig af tiltaltes egne forventninger til algoritmerne, men tilliden kan ligeledes styrkes på en anden måde. Transparensen behøver ikke kun være i forhold til, hvordan algoritmerne virker, men også hvordan de bliver brugt. Dommeren kan forklare, hvordan algoritmen i sidste ende spiller en supplerende rolle i den juridiske proces, og på den måde vise, hvordan algoritmen bidrager til dommerens endelige afgørelse (Ryberg, 2020), og herved styrke befolkningens tillid. Dog er indvendingen mod dette sidste argument lige netop, hvad vi har set på i det ovenstående om psykologisk opacitet.

På baggrund af overvejelserne om, at udgangspunktet for ens forventninger, i forhold til algoritmen, er vigtige for opbyggelsen af tillid, kan det siges, at indvendingen om, at dommere også er

intransparente, bliver styrket. Dette er grundet, at det ikke er muligt, at forklare hele beslutningsprocessen som dommeren foretager sig.

Hvis vi antager at lukkede algoritmer er mere præcise, så vil lukkede algoritmer være at foretrække, frem for åbne, såfremt præcision vægtes højere end transparens. Ved lukkede algoritmer vil der kunne opnås så præcis en risikovurdering som muligt. Og hvis der kan opnås mere præcise risikovurderinger, så har dommerne de bedste forudsætninger for at nå den mest retfærdige dom. Fra dette synspunkt burde lukkede algoritmer ikke medføre mistillid i befolkningen.

5.1.4 Tredje indvending: Det er ikke korrekt at lukkede algoritmer vil skabe mere mistillid i befolkningen, da vi også benytter lukkede algoritmer andre steder i samfundet og dette skaber ikke mistillid

Der er endnu ikke foretaget omfattende og fyldestgørende empiriske undersøgelser af hvorvidt manglende transparens leder til mistillid i befolkningen. Derfor kan det ikke siges klart, at dette skulle være tilfældet at intransparente algoritmer skaber mistillid. Samtidig, hvis dette skulle være sandt, så ville det være det samme som at sige, at vi ikke kan godtage lukkede algoritmer andre steder i samfundet, da vi ikke forstår deres virken. Vi kan derfor ikke argumentere for en hovedløs stræben efter transparens da, “we should [not] give up completely on pursuing transparency, but that we need to be clearer about what we are seeking” (Seymour, 2018: 6 i pdf). Det er relevant at få defineret, hvorfor denne transparens er nødvendig for at skabe tillid, og præcist, hvad denne transparens skal omfatte. For eksempel ved den almene befolkningen ikke nødvendigvis, hvordan algoritmerne fungerer i telefoner, men alligevel benytter vi os hele tiden af dem og med en god portion tillid. Selv hvis det siges at den empiriske antagelse er korrekt; intransparente algoritmer skaber mistillid, så er det ikke nødvendigvis tilstrækkeligt til at kunne sige, det er forkert at bruge intransparente algoritmer, da der kan være andre opvejende faktorer.

5.1.5 Opsamling

Vi har i dette afsnit taget udgangspunkt i argumentet; manglende transparens skaber mistillid i befolkningen. Dette argument synes vigtigt at tage til overvejelse, da samarbejdet mellem befolkning og retssystemet er vigtigt at opretholde, samt at dette fungerer mest optimalt.

Vi har haft følgende indvendinger mod argumentet;

Første indvending: Transparens af koden er irrelevant, da befolkningen alligevel ikke forstår den. Selvom den almene borger havde adgang til kildekoden bag de anvendte algoritmer, er det de færreste i befolkningen, som har forstand på kodning og dets sprog. Altså ville transparens omkring kildekoden ikke understøtte den almene borger i at forstå algoritmerne og dermed ikke skabe større tillid.

Anden indvending: Dommere kan også betragtes som intransparente, så hvis intransparente algoritmer skaber mistillid, så må det samme kunne siges om dommere. Her diskuterer og argumenterer vi for at en dommer, i sig selv, kan anses for at være intransparent, derfor er det ikke klart at en intransparent algoritme skulle øge mistilliden til retssystemet blandt befolkningen.

Tredje indvending: Det er ikke korrekt at lukkede algoritmer vil skabe mere mistillid hos befolkningen, da vi også benytter lukkede algoritmer andre steder i samfundet og dette skaber ikke mistillid.

Der er ikke lavet nok empiriske undersøgelser til at understøtte argumentet; intransparens skaber mistillid i befolkningen - i forbindelse med algoritmer såsom COMPAS og derfor kan det ikke siges at være et tilstrækkeligt argument for ikke at benytte lukkede algoritmer som COMPAS. Endvidere vil transparens af algoritmens kodning, ikke nødvendigvis skabe mere tillid hos den almene befolkning, da befolkningen ikke forstår kodningen. Det er altså ikke tilstrækkeligt at sige at, fordi der ikke eksisterer empirisk data, der viser at intransparente algoritmer skaber mistillid, så må der ikke benyttes lukkede algoritmer i et retssystem.

5.1.6 Delkonklusion

På baggrund af de ovenstående tre indvendinger, vi lige har taget udgangspunkt i, kan vi drage følgende konklusioner:

Vi er kommet frem til at argumentet; intransparens skaber mistillid i befolkning, ikke er et tilstrækkeligt argument til at sige at lukkede algoritmer ikke må benyttes. Endvidere kan der argumenteres for, at en dommer ligeledes kan anses for at være intransparent, og at dommeren alene, ikke er at foretrække, eller at mistilliden ikke vil være stigende, ved brugen af en lukket algoritme. Ved disse tre indvendinger, mener vi, at der på nuværende tidspunkt ikke er tilstrækkeligt med belæg for, at kunne føre argumentet.

5.2.0 Andet argument: Intransparens vanskeliggør en appeldomstols vurdering af en domfældelse, da begrundelserne bag vurderingen er ukendte

En proprietær risikovurderings algoritme som COMPAS betyder at begrundelserne bag algoritmens vurdering er ukendte, hvilket gør det sværere for en højere retsinstant at vurdere præcisionen om en sag, eksempelvis i en situation hvor tiltalte og forsvarer har valgt at appellere sagen. Studier peger nemlig i retning af at begrundelsen bag afsigelsen af en dom, gør det langt lettere for en højere retsinstant at revurdere sagen. Dette betyder at i sager som er afgjort af en dommer, har en appeldomstol lettere ved at vurdere præcisionen samt validiteten i den pågældende dom, ved at kunne spørge direkte ind til bagvedliggende begrundelser for vurderingen (Ryberg, 2020).

5.2.1 Diskussion af argumentet

Vi vil nu gå ind og se på to påstande for at argumentet; Intransparens vanskeliggør en appeldomstols vurdering af en domfældelse, da begrundelserne bag vurderingerne er ukendte, er tilstrækkeligt. Vi vil ligeledes se på en indvending og derved vurdere hvorvidt præmisserne for argumentet er fyldestgørende.

Kan det siges at være sandt, at intransparens omkring en domfældelse, gør det svære at vurdere dommen af en højere retsinstans i en eventuel ankesag?

5.2.2 Første påstand: Transparens øger præcisionen af dommen

Studier viser at når en dommer skal begrunde, hvordan han/hun er kommet frem til den pågældende dom, øger dette præcisionen af dommen.

En dommer er nødt til at være ekstra påpasselig med at begrundelserne for den givne dom bunder i fakta, desuden kan dette samtidig lede til mindre bias i en domfældelse, da lignende sager bør bedømmes på lige fod og ud fra samme juridiske grundlag. Ydermere lader det til, at det er lettere at vurdere præcisionen af en dom, hvis appeldomstolen kan vurdere belægget for den fremsatte dom (Ryberg, 2020). Altså gør transparens det langt lettere at vurdere, hvorvidt en domfældelse er baseret på et præcist grundlag.

Det kan potentielt være langt lettere at vurdere, hvorvidt en dom er præcis, hvis begrundelserne bagved er kendt og baseret på fakta, i modsætning til blot at få tildelt en dom og ikke vide, hvordan domstolen er kommet frem til den givne afgørelse. Altså kan der argumenteres for, at intransparens kan vanskeliggøre en appeldomstols domsfældelse, dertil kan det vise sig underordnet, hvorvidt intransparensen skyldes en proprietær algoritme eller mangel på begrundelse fra en dommer.

Ikke desto mindre er det lettere at efterspørge begrundelser fra en dommer end fra en proprietær algoritme. Dette giver appeldomstolen en lettere mulighed for validering af en given beslutning. Summa summarum er at intransparens vanskeliggør valideringen af en risikovurdering, uanset intransparensens oprindelse.

5.2.3 Anden påstand: Der er flere årsager til at intransparente modeller ikke bør anvendes, end at det nødvendigvis går ud over en appeldomstols vurdering af en sag, blandt andet tiltaltes ret til indsigt i begrundelser bag dommen

Det virker ligetil at argumentere for, at en domfældelse, så vidt det er muligt, bør være transparent, eftersom dette samtidig skulle kunne lede til en mere præcis og mindre bias dom – I hvert fald når der er tale om en domfældelse fra en dommer (jævnfør påstanden ovenfor).

Gør selvsamme sig imidlertid ligeledes gældende for algoritmer? Som det fremgår i afsnittet, af Rudins argumentation mod brugen af black box ML modeller, kan der være en lille forskel i præcisionen af komplicerede black box modeller overfor mere simple transparente modeller. Rudin argumenterer dog for at den transparente model i sidste ende har fordelen ved, at præcisionen af denne er lettere at validere (Rudin, 2019).

Er der hermed belæg for at ML modellerne, der bruges i domfældelse, ikke bør være intransparente? Ja, men ikke nødvendigvis fordi dette betyder at, en appeldomstol vil have sværere ved at vurdere præcisionen af en sag, men nærmere fordi; 1. den tiltalte har ret til at kende begrundelserne bag sin dom, 2. begrundelser lader til, at føre til en mere præcis og mindre bias dom, 3. ved brug af algoritmer er det lettere at validere en transparent model frem for en intransparent model, 4. begrundelser fører i sidste ende til en mere præcis vurdering.

Det kan derfor være at foretrække at undgå intransparens ved en domfældelse. Samtidig ville transparens i algoritmerne også betyde at tiltalte, samt forsvarer, ikke blot var nødsaget til blot at acceptere algoritmens risikovurdering eller validitet (Carlson, 2017). Eftersom transparens kort og godt fører til et mere præcist output, samt en lettere validering af, og måske endda en mere fair risikovurdering. Hvilket er hvad der ønskes af retssystemet (Carlson, 2017). Dette er dog ikke det samme, som at algoritmer ikke bør anvendes ved en domfældelse, men transparente modeller vil klart være at foretrække. Derfor kan der argumenteres for, med denne påstand, at intransparens i algoritmerne kan være med til, at vanskeliggøre en appeldomstols afgørelse, da begrundelserne bag sagen ellers vil være ukendt.

5.2.4 Første indvending: Intransparente algoritmer vanskeliggør ikke en appeldomstols vurdering af en domsfældelse, selvom begrundelserne bag er ukendte

Vi vil nu fremføre en indvending mod at intransparente algoritmer skulle vanskeliggøre en appeldomstols vurdering af en domsfældelse.

Indvendingen lyder: Det vil ikke være sværere for højere retsinstanser at vurdere en sag, hvis der er blevet anvendt risikovurderinger fra intransparente algoritmer.

Det er svært at se at det skulle være tilfældet, at intransparente algoritmer gør denne proces mere besværlig. Med den begrundelse, at en højere retsinstans går ind og ser på sagens detaljer, hvor der for eksempel tages højde for algoritmens risikovurdering og bruger den som grundlag, og hermed ser på om algoritmen i en juridisk sammenhæng er benyttet korrekt. Det er, i og for sig irrelevant for dommeren at se på, om risikovurderingen er udarbejdet korrekt, da det er op til dommeren selv at afgøre, hvordan risikovurderingen skal anvendes. Endvidere har vi, som nævnt ovenfor, argumenteret for, at intransparente algoritmer kan anses for at arbejde mere præcist og korrekt, og derfor vil det ikke være nødvendigt for dommeren at se, hvordan algoritmen er kommet frem til sit resultat. Er det dermed nok at sige, at det ikke er sværere for højere retsinstanser at vurdere en sag, hvis de anvender risikovurderinger fra intransparente algoritmer, til at det er acceptabelt at anvende intransparente algoritmer? Det er på sin vis i orden at anvende intransparente algoritmer, hvis de kommer med en præcis risikovurdering. Dommeren har alligevel mere fokus på en given sags detaljer fremfor, hvordan algoritmen er kommet frem til sit resultat.

5.2.5 Opsamling

Vi har i dette afsnit taget udgangspunkt i argumentet; intransparens vanskeliggør en appeldomstols vurdering af en domsfældelse, da begrundelserne bag vurderingen er ukendte. Her har vi diskuteret både for og imod argumentets hold ved to påstande samt en indvending.

Første påstand: Studier viser at transparens i form af, at skulle begrunde en afgørelse, øger præcisionen af dommen - der kan argumenteres for at, begrundelser gør en dom langt lettere at vurdere præcisionen af, hvis appeldomstolen kan vurdere belægget af den fremsatte dom. Hermed kan

transparens altså være en afgørende faktor i at en given domfældelse betragtes som værende på et præcist grundlag. Intransparens vanskeliggør, uanset dens oprindelse, valideringen af en dom.

Anden påstand: Der er flere årsager til at intransparente modeller ikke bør anvendes, end at det nødvendigvis går ud over en appeldomstols vurdering af en sag, blandt andet tiltaltes ret til indsigt i begrundelser bag dommen - det er helt essentielt i et fair retssystem at den tiltalte har ret til at kende grundlaget bag sin dom, samtidig lader det til at begrundelser fører til en mere præcis, lettere at validere og måske også en mere fair domfældelse. Derfor kunne transparens være at foretrække ved en domfældelse.

Første indvending: Det vil ikke være sværere for højere retsinstanser at vurdere en sag, hvis der er blevet anvendt risikovurderinger fra intransparente algoritmer. Vi har i denne indvending set at det ikke burde spille en stor rolle overfor dommerens beslutning, om algoritmen er åben eller lukket. Hvis en algoritme giver en mere præcis risikovurdering, hvis den er lukket, så er det at foretrække og dette vil ikke spille ind på en højere retsinstans vurdering af en appelleret dom.

5.2.6 Delkonklusion

På baggrund af de to påstande og ene indvending vi lige har taget udgangspunkt i, kan der konkluderes følgende: Hvorvidt intransparens bør undgås ved en domfældelse afhænger af, hvilke faktorer som vægtes højest i afsigelsen af en dom. Vægtes præcisionen af algoritmens vurdering højere end for eksempel den tiltaltes mulighed for indsigt i begrundelserne for hans/hendes dom, så ville en vurdering foretaget af en præcis, men intransparent algoritme være at foretrække. Hvorimod hvis faktorer som muligheden for validering af dommen eller den tiltaltes indsigt i egen dom vægtes højest, så er en transparent algoritme at foretrække.

5.3.0 Tredje argument: Transparens skaber oplevelsen af en mere fair behandling hos den tiltalte

Det tredje argument vi har taget udgangspunkt i, lyder således: Transparens skaber oplevelsen af en mere fair behandling hos den tiltalte.

Ved dette argument er det vigtigt at forstå at der er forskellige betydninger af ordet fair, og hvornår tiltalte får oplevelsen af fair behandling. I forbindelse med risikovurderings algoritmer finder vi to centrale betydninger af oplevelsen af fair behandling i forlængelse af transparens.

Den første er '*process transparency*' (Seymour, 2018: 2), som betyder, hvor meget vi forstår af den interne tilstand i algoritmen. Jo mere transparent algoritmens kildekode er, og jo mere vi forstår algoritmens opbygning, dets større en oplevelse af fair behandling vil en tiltalt føle.

Dernæst er der '*outcome transparency*' (Seymour, 2018: 2), som er; hvor meget vi forstår valgene, som er lavet af algoritmen med hensyn til, hvordan risikovurderingen er kommet frem til sit resultat (Seymour, 2018). Heri gælder det samme, at hvis en tiltalt oplever at der er fair behandling i retsprocessen, fordi det er tydeligt, hvordan algoritmen bidrager til dommerens beslutning, så er der opnået den mest optimale fair behandling.

Som nævnt tidligere i COMPAS' intransparens-afsnittet, er der forskellige mulige årsager til algoritmens intransparens og spørgsmålet og argumenterne til, hvorfor de er det, åbner derfor også op for en diskussion, om en større åbenhed ved algoritmen, vil skabe en mere fair behandling og en tilfredsstillende retfærdighedsfølelse.

5.3.1 Diskussion af argumentet

I dette afsnit ønsker vi at tage udgangspunkt i tre indvendinger til argumentet: Transparens skaber oplevelsen af en mere fair behandling.

5.3.2 Første indvending: Transparens i en algoritme skaber ikke oplevelsen af fair behandling, da det netop kan vises at algoritmen ikke er fuldkommen objektiv

Selvom en menneskelig dommer skal fremføre sig så objektiv som muligt, er det en af de menneskelige beskaffenheder, aldrig at kunne være helt objektiv. Kan en algoritme være fuldkommen objektiv? Der er et tvetydigt svar, da en algoritme på den ene siden ikke har en underbevidsthed eller er selvtænkende nok, til at den kan danne sig meninger og holdninger om hverken hudfarve, klasse eller uddannelse. Men at den derimod er lavet af folk, der bevidst vælger algoritmes fokus-punkter og hvad den skal dømme ud fra. Det er altså en objektiv algoritme, men programmeret af biased programmører. COMPAS er lavet til at skulle gennemskue, hvor stor tilbøjelighed den kriminelle har for tilbagefald til kriminalitet, altså har skaberen af programmet, dannet sig tanker om, hvad der er skyld i, at nogen har større tilbøjelighed til at begå mere kriminalitet. Her lægger COMPAS blandt andet meget fokus på, hvilken social klasse den kriminelle kommer fra og hvor tiltalte er bosat henne, eller hvilken type miljø. Hvorimod en dommer her modsat, måske ikke nødvendigvis kan se et problem ved, forøgede chancer, for tilbagefald på baggrund af, hvor tiltalte er bosat. Er dette nok til at sige at mere intransparente algoritmer ikke kan benyttes fordi de ikke skaber en oplevelse af en fair behandling? Hvis der var transparens i forhold til, hvordan en algoritme såsom COMPAS vægtede inputs, og det dermed kunne vurderes, hvorvidt algoritmen er diskriminerende eller på anden vis unfair justeret, så ville transparens skabe en følelse af afmagt, da det netop opdages at algoritmen er diskriminerende. Modsat kan et svar til denne indvending være; at intransparente algoritmer godt kan skabe en oplevelse af fair behandling selvom de er intransparente. Da det kan tænkes at en algoritme er mere præcis, hvis den er intransparent, vil det i sidste ende skabe en mere fair behandling at straffen bliver udmålt så præcist som muligt.

5.3.3 Anden indvending: En intransparent algoritme kan godt skabe en oplevelse af fair behandling

Når diskussionen om transparens eller ej, hviler på spørgsmålet om en oplevelse af 'fairness' for den tiltalte, skal der overvejes, hvad der ligger til grunde for at den tiltalte oplever at blive behandlet retfærdigt. Her kan der argumenteres for, at en tiltalt vil være mere tilbøjelig til at opleve sin

retssag som værende fair, hvis han/hun føler at være blevet hørt, “an important aspect of this experience relates to the participant’s ability to express his or her view of the case [...]” (Ryberg, 2020: 8). Hvis oplevelsen lever op til dette ønske, kan det antages at den tiltalte vil have nemmere ved at acceptere sin givne dom. Endvidere argumenterer Tyler for, at kriminelle, der oplever at de er blevet dømt ‘fair’ vil have mindre tilbøjelighed til fremtidig kriminalitet (Tyler hos Ryberg, 2020). Dette er grundet en accept af hele retssystemets virken og dermed ikke at have oplevelsen af, ikke at være blevet accepteret i systemet.

Her bliver Loomis’ retssag relevant at tage til overvejelse (Jævnfør Loomis afsnittet). Loomis giver stærkt udtryk for, at han ikke oplever at være blevet behandlet fair, grundet dommerens brug af COMPAS-algoritmen. En af årsagerne til dette er, en manglende transparens af COMPAS-algorithmens virken. Dog kan der argumenteres for, at hvis COMPAS-algoritmen er den mest præcise algoritme, så kan en intransparent algoritme godt skabe en oplevelse af en mere fair behandling, da alternativet ville være en åben algoritme, som ikke nødvendigvis gav en præcis risikovurdering. Endvidere skal det tages til overvejelse, hvorvidt dette alternativ havde givet Loomis en bedre oplevelse af fairness. Altså; lad os antage at Loomis ville være mere tilfreds, hvis domsafsigelsen var afgivet af en dommer, der ikke benyttede risikovurderinger: Ville Loomis’ på baggrund af dette føle sig mere retfærdigt behandlet, eller gav algoritmen blot Loomis en mulighed for at protestere mod sin dom? Denne påstand, er der intet konkret svar på, men grundet dens relevans, har vi anset det som værende værd at overveje.

Som vi i tidligere afsnit har diskuteret, er vi kommet frem til, at en intransparent algoritme stadig kan være at foretrække, da denne muligvis vil komme med mere præcise vurderinger, og hermed kan betragtes som det mest fair alternativ. Dog er det ikke sikkert at et individ som bliver vurderet af en sådan algoritme, oplever dette som fair og blot er villig til at acceptere udsagnet om, at den intransparente algoritme er mere præcis, og dermed mere fair, end den transparente. Her må der derfor overvejes, om det er et spørgsmål om tilvænning. Lad os tage et eksempel med computere. Vi anser computere som et hjælpemiddel og stoler så meget på dem, at vi lader dem hjælpe os med eksamener, økonomiske og personfølsomme data. Dog var computeren i starten anset, som værende for kompleks til at kunne forstå den, og derfor ikke nødvendigvis en god ting, men derimod noget, der ville ændre vores samfund i en negativ retning. Computeren er, den dag i dag, dog meget mere kompleks end da den kom til, og dog er der kun et meget lille fåtal, som er imod computere i dag. Med dette mener vi at behovet for en decideret transparent algoritme muligvis

vil kunne ændre sig i fremtiden, da efterspørgslen på mere præcise og komplekse værktøjer er i højere kurs.

5.3.4 Tredje indvending: Alt afhænger ikke af algoritmens transparens

Vi vil nu se på en indvending mod at en transparent algoritme vil skabe en større oplevelse af at blive fair behandlet.

En transparent algoritme vil ikke nødvendigvis bidrage til en oplevelse af en mere fair retssag, fordi når der er tale om en algoritme i retsvæsenet, så er der ikke tale om en fuldautomatiseret algoritme, der bestemmer dommen på egen hånd, men derimod et hjælpевærktøj til en mulig dom. Beslutningen ligger i sidste ende hos dommeren, derfor skal algoritmen ikke anses som eneste årsag til den pågældende dom, men blot som et værktøj. Algoritmen er blot et supplement som dommeren aktivt kan vælge at benytte sig af i domsafsigelsen, og netop fordi det er op til dommeren at kunne vælge, at inddrage risikovurderingen fra algoritmen, så giver dette mulighed for at forklare, hvordan algoritmen bliver brugt i sammenspil med andre overvejelser, der fører til den endelige dom. Dette vil muligvis kunne skabe en øget oplevelse af en fair behandling. Det kan i denne sammenhæng foretrækkes at algoritmen er så præcis som mulig, hvilket højst sandsynligt indebærer at algoritmen er intransparent, og dermed efterlades dommeren med de bedste forudsætninger for at nå frem til den mest præcise og fair dom. Det er selvfølgelig stadig muligt at den tiltalte har oplevelsen af at være blevet unfair behandlet.

5.3.5 Opsummering

Vi har nu set på tre indvendinger imod at tredje argument skulle være sandt. Vi vil i det følgende opsummere og se på, hvad de tre indvendinger betyder for argumentet:

Første indvending:

Selvom en menneskelig dommer skal fremføre sig så objektiv som muligt, er det en af de menneskelige beskaffenheder aldrig at kunne være helt objektiv.

Vi har set på indvendingen om at den menneskelige dommer altid vil være biased i form af cementerede menneskelige karakteristika, der gør at dommeren aldrig kan forholde sig helt objektivt til en pågældende sag. Hvis en dommer skulle stole blindt på en intransparent algoritme, skrevet af programmører der selv besidder nogle forudindtagelser til algoritmens kildekode. Her kan det blive problematisk da programmørens egne bias, som kan være videregivet til algoritmen, ikke kan udpeges. Helt konkret ville dette gøre sig gældende for, hvordan en algoritme såsom COMPAS-algoritmen, ville være mere fair, hvis den var transparent i henhold til dens vægtning i kildekoden. Derved ville det være lettere at se om algoritmen var diskriminerende og dermed kunne komme frem til unfair risikovurderinger.

Modsat kan det siges; en intransparent algoritme er at foretrække, hvis denne netop kommer frem til mere præcise risikovurderinger, da dette ville skabe større oplevelse af fair behandling.

Spørgsmålet er, om disse indvendinger nok til at sige, at en transparent algoritme skaber en oplevelse af en mere fair behandling?

Anden indvending:

En intransparent algoritme kan godt skabe en oplevelse af fair behandling. I anden indvending har vi set at algoritmer godt kan være intransparente og samtidig skabe en oplevelse af fair behandling, så længe en intransparent algoritme er det mest præcise alternativ overfor en transparent algoritme. Hvis vi antager at en intransparent algoritme kan skabe en mere præcis risikovurdering, så vil en dom i sidste ende være mere fair, hvis dommeren netop benytter sig af en intransparent algoritmes risikovurdering. I sidste ende kan vi indvende at hovedargumentet; transparens skaber oplevelsen af en mere fair behandling hos den tiltalte, ikke nødvendigvis kan siges at være sandt.

Tredje indvending:

Den tredje indvending; alt afhænger ikke af algoritmens transparens, er styrket af, at algoritmer fungerer som værktøjer, da dommere tager den endelige beslutning. Derfor kan der argumenteres for at anvende den mest eksakte algoritme, til at hjælpe dommeren, med at nå den mest præcise dom. Det betyder at intransparente algoritmer kan være en fordel over transparente algoritmer i at nå den mest præcise og dermed mest fair dom. Det er ikke klart at den tiltales oplevelse af at være blevet behandlet fair vil ændre sig uanset om der anvendes transparente eller intransparente algoritmer, da der ingen præcise undersøgelser er lavet til at underbygge denne hypotese. Det betyder

at indvendingen svækker argumentet om at transparens er nødvendigt for oplevelsen af en fair behandling.

5.3.6 Delkonklusion

I dette afsnit argumenterer vi både for og imod, at transparens skaber en større oplevelse af fair behandling. Vi ser en problematik i en intransparent algoritme, da denne kan indeholde en programmørs egne bias, der ikke kan udpeges, da algoritmen fungerer som en black box model. Dog er den menneskelige dommer aldrig selv helt objektiv, hvor en algoritme som hjælpende værktøj, kan være en positiv tilføjelse. Det vil være nemmere for en tiltalt, at acceptere sin dom, hvis han/hun har en oplevelse af fair behandling og dermed have tillid til retssystemets virken. Afslutningsvist kommer vi med indvendingen, at alt ikke er afhængig af algoritmens transparens, en algoritme skal ikke afsige dommen, men være en tilføjelse i form af risikovurderingen og en hjælp til dommerens beslutningstagen. Her kan der argumenteres for, at det på disse præmisser, er det en fordel ved benyttelsen af den intransparente algoritme, da denne, som vi nævner i tidligere afsnit, vil kunne anses for at arbejde mere præcist. Med de tre indvendinger vi har opstillet ovenfor, kan vi konkludere at, hvorvidt transparens fører til oplevelsen af en mere fair behandling, afhænger af hvilke faktorer som anses for værende de vigtigste i afgivelsen af en fair dom. Er det vigtigste den tiltaltes indsigt i begrundelser bag dommen, måske på bekostning af dommens præcision, eller er præcisionen det vigtigste, på bekostning af transparens i dommens begrundelser?

5.4.0 Fjerde argument: Transparens gør det lettere at opdage og rette fejl i algoritmen

Det fjerde argument vi vil behandle lyder: En intransparent algoritme er problematisk, da det ikke er muligt for udefrakommende eksperter og researchers at evaluere og teste algoritmens mulige fejl, mangler og bias.

5.4.1 Diskussion af argumentet

I denne del af diskussionen, vil vi nu diskutere en påstand for og to indvendinger mod argumentet: Transparens gør det lettere at opdage og rette fejl i algoritmen.

5.4.2 Første påstand: Ekspertter skal have adgang til algoritmens kodning.

Er det rigtigt at der ikke kan detekteres fejl på en intransparent algoritme? Er det algoritmens kodning der er manglende transparens på? Eller er det en manglende forståelse for algoritmens vægtning af input, der søges ved transparens? For at svare på disse spørgsmål kan vi bruge Loomis casen som eksempel (jævnfør Loomis afsnittet):

Dommeren Shirley Abrahamson argumenterer for at COMPAS skal være mere gennemsigtig, dette med tilbageblik på Loomis casen, hun påstår at dommere mangler en basal forståelse for algoritmen og dette er yderst problematisk. Hvorfor det kunne give mening at have eksperter til at bygge bro mellem algoritmens kompleksitet og dommerens forståelse af denne. Påstanden støttes op om af lovkyndige eksperter, samt tekniske eksperter; bekymringerne/problematikkerne forklares i manglen på netop en forståelse af algoritmernes inputs, hvordan disse inputs vejes af algoritmen og hvilke specifikke faktorer, som kan være med til at øge en risiko for usynlig diskrimination på baggrund af race og økonomisk status. Der argumenteres for at disse bekymringer, er grundet en manglende transparens, som opstår ved en lukket algoritme. Der er manglende information om, hvilke forudindtagelser eller skjulte bias, som kan have fundet vejen til algoritmen fra dens udviklere (Kehl, Guo, Kessler, 2017).

Endvidere kan der argumenteres for at løsningen vil være en transparent algoritme, hvor udefrakommende eksperter vil have muligheden for at evaluere og teste algoritmen for mulige fejl, mangler og bias.

Ønsket bliver altså at kunne få bedre indblik i, hvordan algoritmen vejer sine inputs og at dette ikke kan gøres uden en transparent algoritme.

Der kan udvises en vis forståelse for, hvorfor COMPAS er en lukket algoritme, da den er skabt af en privat profit virksomhed, som klart ønsker at beskytte deres produkt mod konkurrenter, hvilket lovgivningen ligeledes støtter op om. Dog mener de at der er fordel ved at give akademikere, regeringen samt andre fagfolk adgang til at vurdere en algoritme som COMPAS, da disse har incitament for at sikre en større kontrol og overvågning af algoritmens virken. Disse arbejder ikke med profit som motivator (Kehl, Guo, Kessler, 2017). Endvidere er det nødvendigt at sikre algoritmens score er fair og præcis - at teste algoritmen, er at sikre dette mål. Ved transparens og åben adgang for akademikere kan der kommenteres på systemets virken og sikre at algoritmen arbejder inden for befolkningens værdier frem for virksomhedens (Kehl, Guo, Kessler, 2017).

Derfor kan det ønskes, som vi også ser i den forudgående debat, at eksperter får adgang til algoritmen og hermed kan få indblik i, hvordan denne vejer de inputs algoritmen, arbejder med. Dette vil altså kræve adgang til selve kodningen af algoritmen.

5.4.3 Første indvending: Kodningen er for kompleks, så selv fagpersoner ikke forstår algoritmen.

Den første indvending lyder: Kodningen er for kompleks så selv fagpersoner ikke forstår algoritmen. Der er ingen tvivl om at transparens, vil give en større mulighed for at fejl tjekke algoritmens virken, der kan hermed argumenteres for at det er en nødvendighed, at give udefrakommende eksperter adgang til systemet for at kunne modarbejde mulig bagvedliggende bias i kodningen og at dette vil være med til at kunne opretholde befolkningens interesse og værdier. Det giver derfor mening at udefrakommende fagfolk, akademikere etc. kan være en hjælpende faktor, da de ikke nødvendigvis har noget at tabe ved at finde fejl i kodningen, deres ry sættes ikke på spil, og hvorfor skulle de ikke have befolkningens interesse i fokus?

Spørgsmålet er så, om problemet overhovedet bunder i manglende transparens, eller om de diskriminative bias, som kan opfanges, i sig selv, nærmere bunder i et mere generelt samfundsproblem? Det må antages, at hvis biases skyldes input i algoritmen, så vil det at algoritmen er åben, ikke ændre på eksisterende bias, da disse er iboende i algoritmen. Dette vil transparens af algoritmens kodning ikke kunne rette synderligt meget op på - dog, hvis dette er tilfældet, kan det ikke siges at være et problem at algoritmen er tilgængelig for udefrakommende eksperter. Hvis problematikken

derimod findes i kodningen, kan transparens gøre forskellen og derfor kan vi ikke se, hvorfor algoritmer ikke skulle være transparente.

Det der egentlig ønskes, er en mere tilstrækkelig forståelse for hvordan, algoritmen vejer sine inputs, dette kan dog kun opnås ved en åben adgang til algoritmens kode - det er her fagfolk og eksperter vil kunne finde svarene.

Selvom man argumenterer for at transparens kan gøre det lettere at rette fejl i kodningen, så kan det indvendes, at adgang til algoritmens kode ikke hjælper, da kodningen er så kompleks at selv eksperter ikke nødvendigvis ville have forståelse for, hvordan algoritmen arbejder.

Er dette nok til at kunne sige, at intransparente algoritmer ikke bør benyttes i retssager? Nej for der kan stadig indvendes at det er mere optimalt at benytte intransparente algoritmer, der giver et bedre resultat, end ved åbne algoritmer, der ikke er lige så præcise. Så indvendingen gør ikke nødvendigvis, at vi kan afgøre eller retfærdiggøre brugen af intransparente algoritmer.

5.4.4 Anden indvending: Risikovurderings algoritmer optimerer sig selv

Visse machine learning algoritmer er programmeret til at finde den bedste og mest optimale sammenhæng mellem input og output data, samt at den hele tiden forsøger at optimere relationen mellem datasættene. Derfor kan det tænkes at en programmør vil opleve, at algoritmen har ændret sig fra den første udgave af kodningen. Det betyder at når der bedes om transparens og åbenhed overfor, hvordan algoritmen er justeret, så vil en programmør, der åbner en algoritme, som har optimeret sin proces, kigge ind i en algoritme, der ikke ligner udgangspunktet. Det vil sige, at hvis han/hun leder efter mulige fejl, eller forståelse for, hvordan algoritmen fungerer, så vil optimeringer, algoritmen selv har lavet, kunne se uforståelige ud for programmøren. Denne indvending gør det let at pege på, at transparens ikke nødvendigvis gør det lettere at finde og rette fejl i kodningen. En anden variant af indvendingen kan også tænkes; hvis der er enighed om en lovgivning, der skal sikre et udgangspunkt for risikovurderings algoritmer, altså hvordan algoritmer skal vægte forskellige input data, eller hvilke data den skal finde irrelevante i forhold til dommen, så må dette udgangspunkt være plastisk og derfor kunne modificeres. For hvis algoritmer hele tiden optimerer processen, så må reglerne også hele tiden optimeres.

Et scenarie kan tænkes; algoritmer definerer og styrer reglerne for udgangspunktet for andre algoritmer, efter menneskelige programmører har defineret reglerne første gang. Det kan derfor ende i et scenarie, hvor algoritmer, der er selvoptimerende og selvjusterende, skal styre reglerne for risikovurderings algoritmer, der også er selvstyrende og selvjusterende, og dermed er vi kommet væk fra al menneskelig involvering i processen. Det er tydeligt at dette kan ende i et glidebaneargument. Som regel er indvendingen mod dette argument, at vi ikke nødvendigvis er nødt til at tænke det værste scenarie om selvstyrende algoritmer, når vi har taget første skridt ved at benytte risikovurderings algoritmer i den juridiske proces. For argumentet i sig selv er lige så meget et gæt på et muligt udfald, som det at sige, at vi ikke nødvendigvis vil gå hele vejen mod det værste scenarie, og sagtens kan stoppe tidligere og justere processen. Dog er det i dette hypotetiske tilfælde allerede ude af menneskelige programmørers hænder, da algoritmerne udvikler sig eksponentielt og konstant optimerer processen. Derfor kan det argumenteres for, at når første skridt tages, så leder det hele vejen til værste scenarie.

Kan det tænkes at transparensen, der skal gøre det lettere at finde og rette fejl, kan forudse denne problematik, og ved netop at være bevidst om, at algoritmerne fungerer på en selvoptimerende og selvjusterende måde, kan vi finde den mest optimale proces og brug af henholdsvis dommere og algoritmer i en given retssag? For hvad er alternativet, hvis vi går den modsatte vej? Det ville muligvis ende med, at hvis vi kommer frem til konklusionen, at transparens alligevel ikke kan bidrage til at finde og rette fejl, så kan vi slet ikke benytte os af algoritmer? Dette argument kan kun bygges op af en sammenligning med status quo på domsafsigelserne fra før risikovurderings algoritmer, som metode, blev taget i brug (Chiao, 2020). Her er hensigten ikke at vurdere om den algoritmiske metode er perfekt transparent, men snarere, hvorvidt den algoritmiske metode er mere eller mindre transparent end status quo, før den algoritmiske metode blev taget i brug, hvor det kun var menneskelige dommere, der afgjorde dommen (Chiao, 2020).

Er dette tilstrækkeligt til at kunne sige, at vi ikke må bruge intransparente algoritmer? Nej, det er ikke tilstrækkeligt til at sige, at vi ikke må benytte os af intransparente algoritmer, da dette stadig vil kunne argumenteres for at være bedre end slet ikke at benytte nogle, her på baggrund af, at vi som mennesker ikke nødvendigvis vurderes til at være de bedste til at forholde os objektivt til en given sag (Se ovenfor; 5.1.3). Vi er selv styret af personlige bias, som ikke nødvendigvis kan udpeges. Derfor vil en intransparent algoritme stadig være at foretrække frem for en dommer alene.

5.4.5 Opsummering

Første påstand; eksperter skal have adgang til algoritmens kodning. Det ønskes at få bedre indblik i, hvordan algoritmer som COMPAS, vejer sine inputs, dette vil være langt lettere at opnå, hvis udefrakommende eksperter kan få adgang til algoritmens kodning. Påstanden her viser, at afhængigt af, hvor der ønskes transparens, vil dette kunne gøre forskellen. Hvis der ønskes adgang til måden COMPAS vejer sine inputs, for at kunne rette op på potentielle fejl, vil transparens af algoritmens kodning være afgørende.

Første indvending: Kodningen er for kompleks så selv fagpersoner ikke forstår algoritmen:

I den første indvending blev der fremlagt, at selvom der bliver givet adgang til en algoritmens kildekode, så er koden så kompleks, at selv fagpersoner ikke nødvendigvis præcis forstår, hvordan algoritmen virker. Det er vigtigt for at standse mulig bias, at udefrakommende eksperter får adgang til systemet for at kunne give modspil og udfordring til algoritmen, for her at standse mulig bias. I denne indvending har vi forsøgt at argumentere for, at transparens ikke nødvendigvis hjælper på at kunne rette fejl i algoritmerne. Vi har vist at indvendingen ikke er nok til at kunne støtte argumentet om, at intransparente algoritmer ikke må benyttes, fordi de gør det sværere at finde og rette fejl.

Anden indvending: Risikovurderings algoritmer optimerer sig selv: Vi har i denne indvending opstillet et hypotetisk scenarie, hvor algoritmer kan forestilles at være selvoptimerende og justerende, her vil det altså ikke nødvendigvis være muligt for programmøren at genkende kodningen på et senere tidspunkt. Afslutningsvist har vi overvejet, hvorvidt det er tilstrækkeligt at sige, at vi ikke må bruge intransparente algoritmer? Her argumenterer vi for, at dette ikke er tilfældet, da det stadig vil være at foretrække, da vi som mennesker, ikke vurderes til at være de bedste til at forholde sig objektivt til en given sag.

5.4.6 Delkonklusion

Det vi har diskuteret er, argumentet om at vi ikke bør bruge intransparente algoritmer, da det ikke er muligt at detektere fejl og mangler i algoritmen.

Først er vi kommet frem til, hvad det egentlig er der ønskes og hvor problematikken ligger. Det der udtrykkes ønske om hos Kehl, Guo og Kessler, er en forståelse for, hvordan algoritmen vejer og måler sine inputs, dette skal gøres ved hjælp af udefrakommende eksperter. Hvis udefrakommende eksperter skal kunne vurdere, hvorvidt algoritmen arbejder på en forsvarlig måde, uden bias og diskriminerende faktorer, skal algoritmen til en vis grad være åben for adgang - det vil altså være nødvendigt med adgang til kodningen for at kunne få det ønskede udfald. Endvidere kan det siges at kodningen i sig selv, er utroligt kompleks og kan dermed være svær at vurdere for selv fagfolk og derfor vil det ikke nødvendigvis være muligt at finde fejl og mangler, selv med en transparent kode.

Afslutningsvist har vi diskuteret hvorvidt det er tilstrækkeligt at sige, at der ikke må benyttes intransparente algoritmer. Her er svaret nej, en intransparent algoritme vil stadig være et foretrukket hjælpemiddel, frem for status quo; en dom baseret udelukkende på dommerens egen intuition.

6.0 Konklusion

Vi har igennem rapporten undersøgt problemstillingen: Er det etisk acceptabelt at benytte intransparente algoritmer til risikovurdering i retssystemet?

I løbet af denne undersøgelse er vi nået frem til fire hovedargumenter som vi har gennemgået systematisk.

Argumenterne lyder som følger:

- Intransparente algoritmer bør ikke anvendes, da det vil skabe mistillid i befolkningen.
- Intransparens vanskeliggør en appeldomstols vurdering af en domfældelse, da begrundelserne bag vurderingen er ukendte.
- Transparens skaber oplevelsen af en mere fair behandling hos den tiltalte.
- Transparens gør det lettere at opdage og rette fejl i algoritmen.

Efter en systematisk gennemgang, er vi kommet frem til at vi kan konkludere følgende:

Det første argumentet vi har diskuteret; intransparente algoritmer bør ikke anvendes, da det vil skabe mistillid i befolkningen, kan ikke siges at være tilstrækkeligt til at kunne sige, at det ikke er

forsvarligt at benytte intransparente algoritmer i retssystemet. Selvom befolkningen skulle få adgang til kodningen, så vil det ikke nødvendigvis skabe en større forståelse af den, og dermed ikke være argument nok til at sige, at der ikke bør benyttes intransparente algoritmer. Ligeledes kan en dommer anses for selv at være intransparent og derfor kan det siges at en domsafgørelse, baseret alene på dommerens intuition, ikke er at foretrække. Det må altså kunne accepteres at der anvendes intransparente algoritmer, hvis det giver en større præcision i domsafsigelserne. Endvidere er der ikke evidens nok til at kunne sige, at en lukket algoritme faktisk vil skabe mere mistillid i befolkningen. Dette er vi kommet frem til på baggrund af, at mistilliden ikke nødvendigvis afhænger af mere eller mindre transparens, men også af andre faktorer som vi vil gennemgå i konklusionen af de følgende argumenter.

I henhold til andet argument i vores diskussion; intransparentens vanskeliggør en appeldomstols vurdering af en domfældelse, da begrundelserne bag vurderingen er ukendte, kan dette heller ikke siges at være incitament nok til at der ikke bør bruges intransparente algoritmer. Her bliver den afgørende faktor, hvad der vægtes højest i afsigelsen af en dom. Vægtes præcisionen af algoritmens vurdering højere end indsigt i begrundelsen for en given dom, så vil en vurdering foretaget af en intransparent algoritme være at foretrække, hvorimod hvis validering og indsigt vægtes højest, så er en transparent algoritme at foretrække.

I det tredje hovedargument kommer vi frem til følgende; det er ikke klart hvorvidt transparens leder til en øget oplevelse af fair behandling. Der kan argumenteres for at transparens skaber en oplevelse af en mere fair dom, fordi begrundelserne for dommen er kendt, og dermed giver tiltalte bedre indsigt i sin dom. Samtidig kan der også argumenteres for det modsatte, da brugen af en intransparent, men mere præcis algoritme bedre kan assistere dommeren i at nå frem til den mest fair dom. Det er dog ikke klart om dette har en betydning for selve oplevelsen af en fair dom. Derfor må det være et spørgsmål om, hvilke faktorer der menes at vægte højest for at opnå en fuldbyrdet oplevelse af fair behandling. Er indvendingerne nok til at sige, at der ikke bør anvendes intransparente algoritmer? Vi kan konkludere, at hvis det er et krav, at det tredje argument skal være gyldigt og til stede i en retsproces, så skaber transparens rent faktisk en større oplevelse af fair behandling, hvis vi har at gøre med algoritmer, der vægter dens input data på en diskriminerende måde. Dog kan det siges ikke at være sandt, hvis vi antager at en intransparent algoritme kan

komme med en mere præcis risikovurdering når den er lukket, også selvom programmørerne ikke nødvendigvis ved, hvordan den virker.

Afslutningsvist har vi diskuteret validiteten af fjerde argument; transparens gør det lettere at opdage og rette fejl i algoritmen. Her er vi først kommet frem til, hvad der egentlig ønskes; dette er en forståelse for, hvordan en algoritme måler og vejer dens inputs. Endvidere skal det være udefrakommende eksperter som efterser algoritmens virken. Hvis dette er målet, skal en algoritme være delvist åben, så eksperterne kan få adgang til kodningen. Vi argumenterer for, at kodningen i sig selv er en utroligt kompleks størrelse og dermed kan den selv for fagfolk være svær at vurdere. Altså vil det ikke nødvendigvis være muligt, at finde fejl og mangler ved en transparent kode.

Ud fra de argumenter vi har behandlet, med hensyn til, hvorvidt det er etisk korrekt at benytte intransparente algoritmer til risikovurdering i et retssystem, finder vi konklusionen tvetydig. Dette skyldes, at vi mener problemstillingen må være til diskussion; vi kan hverken konkludere om det er etisk uacceptabelt eller acceptabelt. Der er forskellige faktorer, der påvirker dette, såsom hvad der i realiteten forlanges af algoritmen og det kræver først og fremmest en mere præcis klargøring af, hvad der skal forventes af den ønskede algoritme, når den benyttes i en stats retssystem, for derefter at gøre det muligt, at konkludere, om den er etisk acceptabel eller ej.

7.0 Litteraturliste

- ⇒ Angwin, Julia, Surya, Mattu, Lauren Kirchner Et al. (2020): *Machine Bias*. ProPublica. [Online] <<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>> [Sidst tilgået, 17.12.20, kl. 15.32]
- ⇒ Carlson, Alyssa. M. (2017): *The Need for Transparency in the Age of Predictive Sentencing Algorithms*. [Online] <<https://ilr.law.uiowa.edu/print/volume-103-issue-1/the-need-for-transparency-in-the-age-of-predictive-sentencing-algorithms/>> [Sidst tilgået, 17.12.20, kl. 10.38]
- ⇒ Corbett-Davies, S. E. P. (2016): *A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear*. Washington Post. [Online] <<https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/>> [Sidst tilgået, 01.12.20, kl. 14.24]
- ⇒ Department of Computer Science (Ukendt): *William Seymour*. [Online] <<https://www.cs.ox.ac.uk/people/william.seymour/>> [Sidst tilgået, 15.12.20, kl. 15:45]
- ⇒ Internet Law & Policy Foundry (Ukendt): *Danielle Kehl*. [Online] <<https://www.ilpfoundry.us/team/danielle-kehl/>> [Sidst tilgået 10.12.20, kl. 15.21]
- ⇒ Kehl, Danielle, Priscilla Guo, and Samuel Kessler (2017): *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing*. University of Harvard. [Online] <https://dash.harvard.edu/bitstream/handle/1/33746041/2017-07_responsiblecommunities_2.pdf?sequence=1&isAllowed=y> [Sidst tilgået, 04-12-20, kl. 13.23]
- ⇒ Kessler, Samuel Joseph (ukendt): *Samuel Joseph Kessler*. [Online] <<http://samueljkessler.com>> [Sidst tilgået 16.12.20, kl. 15.02]
- ⇒ Northpointe inc. (2012): *COMPAS Risk & Need Assessment System*. [Online] <http://www.northpointeinc.com/files/downloads/FAQ_Document.pdf> [Sidst tilgået, 25.11.20, kl. 13.21]
- ⇒ Octomaslaw (2019): *Alyssa M. Carlson*. [Online] <<https://www.octomaslaw.com/professional/alyssa-m-carlson/>> [Sidst tilgået, 28.12.20, kl. 12.15]

- ⇒ Oxford Internet Institute (Ukendt): *Priscilla Guo*. [Online]
<<https://www.oii.ox.ac.uk/people/priscilla-guo/>> [Sidst tilgået, 16.12.20, kl. 14.39]
- ⇒ ProPublica, a (2019): *Julia Angwin*. ProPublica. [Online]
<<https://www.propublica.org/people/julia-angwin>> [Sidst tilgået, 24.11.20, kl. 15.37]
- ⇒ ProPublica, b (2019): *Surya Mattu*. ProPublica. [Online]
<<https://www.propublica.org/people/surya-mattu>> [Sidst tilgået, 24.11.20, kl. 15.37]
- ⇒ ProPublica, c (2020): *Lauren Kirchner*. ProPublica. [Online]
<<https://www.propublica.org/people/lauren-kirchner>> [Sidst tilgået, 24.11.20, kl. 15.37]
- ⇒ Rudin, Cynthia (2019): *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*. [Online] <<https://doi.org/10.1038/s42256-019-0048-x>> [Sidst tilgået, 17.12.20, kl. 12:23]
- ⇒ Ryberg, Jesper (2020): *Chapter 2: Sentencing and Algorithmic Transparency* [Endnu ikke udgivet].
- ⇒ Ryberg, Jesper (Ukendt): *Jesper Ryberg* [Online] <<https://ryberg.wixsite.com/jesper>> [Sidst tilgået, 23.11.20, kl. 16.43]
- ⇒ Seymour, William (2018): *Detecting Bias: Does an Algorithm Have to Be Transparent in Order to Be Fair?* [Online] <http://ceur-ws.org/Vol-2103/paper_1.pdf> [Sidst tilgået, 16.12.20, kl. 10.30]
- ⇒ Springer, A. and S. Whittaker (2018): *'I Had a Solid Theory Before But it's Falling Apart'*:
⇒ *Polarizing Effects of Algorithmic Transparency*. [Online] <arXiv:1811.02163> [Sidst tilgået, 15.12.20, kl. 11.04]
- ⇒ University of Duke (2020): *Cynthia Rudin*. [Online] <<https://users.cs.duke.edu/~cynthia/>> [Sidst tilgået, 24.11.20, kl. 17.35]
- ⇒ University of Toronto (Ukendt): *Vincent Chiao*. [Online] <<https://www.law.utoronto.ca/faculty-staff/full-time-faculty/vincent-chiao>> [Sidst tilgået, 22.11.20, kl. 15.34]

7.1 Bilag

⇒ Angwin, Julia (2020): *Sample-COMPAS-Risk-Assessment-COMPAS - "CORE"*. ProPublica [Online] <<https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html>> [Sidst tilgået, 10-12-20, kl. 16.32]