KI TEKNOLOGI

Kunnskapsoversikt knyttet til forsvarlig bruk av kunstig intelligens i petroleumssektoren

Havindustritilsynet

Rapportnr.: 2024-1519, Rev. 1
Dokumentnr.:
Dato: 19.12.2024

Prosjektnavn:	KI teknologi	DNV Energy Systems
Rapporttittel:	Kunnskapsoversikt knyttet til forsvarlig bruk av kunstig intelligens i petroleumssektoren	Digital Technoloy, Høvik Veritasveien 1, 1322 Høvik
Oppdragsgiver:	Havindustritilsynet, Proff. Olav Hanssens vei 10, 4021 Stavanger	Tel: +47 67579900 945748931
Kontaktperson:	Linn Iren Vestly Bergh	
Dato:	19.12.2024	
Prosjekt nr.:	10434783	
Org. enhet:	Digital Technology, Subsea, Risers and Pipelines	
Rapportnr.:	2024-1519, Rev. 1	
Dokumentnr.:		

Levering av denne rapporten er underlagt bestemmelsene i relevant(e) kontrakt(er):

Oppdragsbeskrivelse:

Havtil har gitt DNV i oppdrag å utarbeide en kunnskapsoversikt som dekker de grunnleggende risikoforholdene ved utvikling og bruk av kunstig intelligens (KI) i petroleumsvirksomheten, spesielt med hensyn til storulykkesrisiko.

Utført av:	Verifisert av:	Godkjent av:
Digitally signed by Christian Markussen Date: 2024.12.20 08:09:11 +01'00' Christian Markussen Global Practice Lead	Agrell, Christian Digitally signed by Agrell, Christian Date: 2024.12.19 23:11:42 +01'00' Christian Agrell Principal Researcher	Jahre-Nilsen, Per Digitally signed by Jahre-Nilsen, Per Date: 2024.12.19 16:07:20 +01'00' Per Jahre-Nilsen Head of section
Meine van der Meulen Digitally signed by Meine van der Meulen Date: 2024.12.20 12:26:40 +01'00' Meine van der Meulen Senior Principal Researcher		
Digitally signed by Kvinnesland, Kenneth Date: 2024.12.19 16:00:02 +01'00' Kenneth Kvinnesland Senior Principal Consultant		
van de Merwe, Koen Digitally signed by van de Merwe, Koen Date: 2024.12.19 16:26:50 +01'00' Koen van de Merwe Principal Reseracher		

Internt i DNV er informasjonen i dette dokumentet klassifisert som:

	Kan dokumentet bli distribuert internt i DNV etter en gitt dato? Nei	Ja
☒ Open ☐ DNV Restricted	--	--
☐ DNV Confidential ☐ DNV Secret	☐	☐

Flere personer som er autorisert til å distribuere dette dokumentet internt i DNV:

Keywords: Kunstig Intelligens, Storulykker, Petroleumssektoren

Rev. no.	Date	Reason for issue	Prepared by	Verified by	Approved by
0	2024-10-25	Første utgave	Christian Markussen Meine van der Meulen Koen van de Merwe Kenneth Kvinnesland	Christian Agrell	Per Jahre-Nilsen
1		Revidert for språk			

Innholdsfortegnelse

1 SAMMENDRAG

Havindustritilsynet (Havtil) har gitt DNV i oppdrag å utarbeide en kunnskapsoversikt som dekker de grunnleggende risikoforholdene ved utvikling og bruk av kunstig intelligens (KI) i petroleumsvirksomheten, spesielt med hensyn til storulykkesrisiko. Dette sammendraget beskriver kort de viktigste observasjonene i rapporten.

Hvor forventes KI innført

DNV forventer at KI vil bli innført i petroleumsvirksomheten i et høyt tempo. I første omgang vil KI vil bli anvendt for å generere forskjellige typer dokumenter og kildekode, og vil på kort sikt bli brukt i forskjellige typer av rådgivende applikasjoner, f.eks. innen driftsoptimalisering og prediktivt vedlikehold. På lengre sikt forventes KI også brukt i kontrollfunksjoner, for eksempel knyttet til løfteoperasjoner og brønnkontroll. Autonome KI-baserte systemer blir i første omgang innført der skadepotensialet er lavt, eksempelvis i undervannsfarkoster og droner.

Risikoer knyttet til bruk av KI

Denne rapporten identifiser en rekke KI-relaterte risikoer som kan føre til at operasjoner havner i en utrygg tilstand. Noen eksempler er: utilstrekkelig trening av algoritmer, dårlig datakvalitet, modellforringelse over tid, og overtilpasning til treningsdata. I tillegg vil også KI-baserte systemer være sårbare for svakheter i programvaredesign, feil i maskinvare, samt villede handlinger som cyber-angrep. Som beskrevet ovenfor forventes KI tatt i bruk i forskjellige typer av systemer og operasjoner, og for de fleste av disse er det en fare for at informasjon generert ved hjelp av KI-baserte systemer, kan føre til at det blir tatt feil operasjonelle beslutninger, noe som igjen kan føre til en ulykke. Av den grunn bruker denne rapporten begrepet «sikkerhetsrelaterte systemer» som et samlebegrep som omfatter sikkerhetssystemer, kontroll og overvåkingssystemer, samt systemer for rådgivning, planlegging, og tilstandsovervåkning.

Bruk av barrierer

Sikkerhet for menneske og miljø forventes også i fremtiden ivaretatt gjennom barrierefilosofien som ligger til grunn for Havtils regelverk. Tankegangen bak barrierefilosofien er at uansett hvor godt en forsøker å få til en sikker og robust løsning, vil feil, fare- og ulykkessituasjoner kunne inntreffe, og da skal barrierer tre i kraft og bidra til å håndtere slike situasjoner. Dersom en KI-basert applikasjon skulle gi resultater som gjør at en operasjon havner i en utrygg tilstand, tilsier denne filosofien bruk av barrierer for å hindre at tilstanden eskalerer til en farlig hendelse. Eksempler på slike barrierer er menneskelig overstyring, bruk av uavhengige kontrollfunksjoner, bruk av uavhengige sikkerhetsfunksjoner som stenger ned prosessen som blir kontrollert, samt bruk av operasjonelle restriksjoner som reduserer risikoen for en hendelse dersom de andre barrierene ikke skulle være effektive.

Samspillet mellom KI og mennesker

For mange typer av operasjoner vil det i dag være mennesker som tar beslutningen om å aktivere barrierer av den typen som er beskrevet ovenfor. Utfordringen med det ligger blant annet i at uønskede tilstander i programvare ikke alltid fører til alarmer. Eksempelvis er det ikke gitt at man får en alarm dersom en KI-algoritme blir utsatt for et operasjonelt scenario den ikke er trent for. I en slik situasjon vil barrierer kun bli aktivert dersom mennesker er i stand til å forstå at noe er galt basert på den totale mengden tilgjengelig informasjon. Av den grunn fokuserer denne rapporten særlig på samspillet mellom KI og mennesker, behovet for et menneskesentrert design av KI-løsninger, samt behovet for å kunne detektere utrygge tilstander på en måte som er uavhengige av systemet som inneholder KI.

Uavhengig menneskelig deteksjon av utrygge tilstander vil i noen tilfeller kunne være mulig gjennom å sammenligne resultater produsert av KI med resultater produser ved hjelp av en alternativ teknolog. Målinger av prosessparametere kan også brukes til å indentifisere utrygge tilstander forårsaket av KI, men vil ikke alene kunne brukes til å detektere alle former for utrygge tilstander. Dette skyldes at felles-feil i programvare eller problemer med datakvalitet vil kunne påvirke både de beslutningsprosessene som foregår i programvare og de som foregår hos operatør negativt. Deteksjon basert på menneskelig observasjon av prosessparametere og annen relatert informasjon, krever også at det er tid nok til å gjøre gode vurderinger, samt dyp innsikt i prosessen som blir kontrollert. Slik innsikt vil som regel være viktigere enn innsikt i hvordan KI fungerer.

Uansett kunnskapsnivå vil det kunne være vanskelig å identifisere avvikssituasjoner manuelt, dersom avvikene fra normal prosess er relativt små, noe som vil kunne være et problem dersom man også operer med små marginer mot utrygge prosesstilstander.

KI vil typisk øke både kapabilitet og kompleksitet for et system. Forskning viser at jo mere kapabelt og komplekst et system blir, jo mindre evne har brukeren til å forstå systemet og dets begrensninger, og til å overvåke det på en pålitelig måte. Utfordringen knyttet til menneskelig deteksjon av utrygge tilstander, tilsier at industrien bør utforske mulighetene for automatisert deteksjon av utrygge tilstander forårsaket av KI.

Faktorer som kan begrense hvor KI kan innføres

For at barrier skal være effektive, må det være mulig å oppdage at en usikker tilstand har oppstått, uavhengig av hva som har forårsaket den – ellers kan man ikke stole på at barrierene blir aktivert ved behov. Hvis programvare på en eller annen måte har forårsaket en usikker tilstand, er det kun mulig å oppdage dette dersom det finnes en deteksjonsmekanisme som er helt uavhengig av denne programvaren. En sikkerhetsfunksjon som aktiveres automatisk når målinger av kritiske prosessparametere overstiger en forhåndsdefinert terskelverdi, er et eksempel på en barriere der slik uavhengig deteksjon er tilgjengelig.

Økt grad av automasjon gjør imidlertid at det at det i mange tilfeller kan være vanskelig å detektere uønskede tilstander på en uavhengig måte uansett hva som har forårsaket tilstanden. Det betyr at petroleumsindustrien beveger seg mot en gråsone der man som i en del andre industrier ser kritiske og komplekse funksjoner som kontinuerlig må virke etter intensjon for at sikkerheten skal være ivaretatt. Som beskrevet ovenfor er utfordringene særlig knyttet til menneskelig deteksjon av uønskede tilstander. Disse utfordringene vil ofte være til stede uavhengig av om KI benyttes eller ikke, men bruk av KI kan forsterke problemet.

På grunn av at barrierefilosofien krever separasjon av kontroll- og sikkerhetsfunksjoner, er kravene til hvordan programvarebaserte kontrollfunksjoner utvikles, verifiseres og valideres ikke særlig strenge. Dette innebærer at terskelen for å bruke kunstig intelligens (KI) i kontrollfunksjoner kan være lavere enn terskelen for å bruke KI i sikkerhetsfunksjoner. Mangelen på mekanismer som kan oppdage usikre tilstander, uavhengig av om systemet bruker KI, forventes likevel å være en begrensende faktor for hvor KI kan innføres på en trygg måte

Standardene som Havtils regelverk refererer til når det gjelder rene sikkerhetsfunksjoner stiller svært strenge krav til hvordan programvare skal utvikles, verifiseres og valideres, og innføring av KI i slike funksjoner vil dermed utløse en betydelig bevisbyrde. Av den grunn er ikke introduksjon av KI i slike funksjoner forventet i nær fremtid. Det kan likevel ikke utelukkes at KI-baserte komponenter kan bli innført på lengre sikt, eksempelvis kan man tenke seg KI-basert aktivering av sikkerhetsfunksjoner kommer som et tillegg, der man i dag kun har menneskelig aktivering.

For noen KI-systemer så vil resultatene som blir produsert ikke være deterministiske. Det betyr at dersom man utfører flere tester med samme inngangsdata, så får man ikke nødvendigvis de samme resultatene, noe som kan vanskeliggjøre kvalifisering, validering og også vedlikehold av programvare som inneholder KI. Dette kan også være en begrensende faktor for hvor KI kan innføres, og industrien bør utforske hvordan denne utfordringen kan løses.

Behov for felles veiledninger for bruk av KI i petroleumsindustrien

Havtils regelverk henviser til en lang rekke standarder og veiledninger, både Norske og internasjonale. Disse velger de fleste aktørene å følge, siden man ellers må demonstrere at alternative tilnærminger er like bra eller bedre. Bruk av felles standarder og veiledninger bidrar til et harmonisert sikkerhetsnivå i petroleumsindustrien. Det finnes imidlertid så langt ingen standarder eller veiledninger som er utarbeidet spesifikt med tanke på sikker bruk av KI innenfor petroleumsindustrien. For å opprettholde et harmonisert sikkerhetsnivå, og for å redusere bevisbyrden for den enkelte aktør, vil det være ønskelig at relevante aktører i industrien går sammen om å utarbeide veiledninger som representerer beste praksis for bruk av forskjellige typer av løsninger som inneholder KI.

Et slikt arbeide bør også kunne gjøre det enklere å møte kravene i EU forordningen om KI. Foreløpig må den enkelte aktør som ønsker å ta i bruk KI selv konkretisere og svare ut kravene i forordningen i egne styringssystemer. Slik operasjonalisering av høynivå krav er vanligvis arbeidskrevende, men arbeidsbyrden på hver enkelt organisasjon bør kunne reduseres dersom industrien går sammen om det.

2 INNLEDNING

2.1 Oppdragsbeskrivelse

Havtil har gitt DNV i oppdrag å utarbeide en kunnskapsoversikt som dekker de grunnleggende risikoforholdene ved utvikling og bruk av kunstig intelligens (KI) i petroleumsindustrien, spesielt med hensyn til storulykkesrisiko. Formålet med studien er å øke kunnskapen om risikoer ved utvikling og bruk av KI i operasjoner som har betydning for sikkerheten på norsk kontinentalsokkel.

Kunnskapsoversikten skal bidra med økt kunnskap om risikoer assosiert med KI-systemer i petroleumssektoren. Den skal utforske hvordan KI kan forbedre både effektivitet og sikkerhet, samtidig som man tar hensyn til de unike risikoene KI introduserer sammenlignet med tradisjonelle IT- og automatiseringssystemer.

Det er nødvendig å benytte oppdatert litteratur og kunnskap om KI-modeller og deres potensielle konsekvenser for sikkerheten i petroleumsindustrien. Dette omfatter forhold som datakvalitet, integritet, trening, re-trening, testing, transparens og forklarbarhet samt implementering. Det inkluderer også metoder og teknikker for å identifisere og håndtere risiko i ulike faser av KI-modellenes livssyklus, og vektlegging av god situasjonsforståelse og kommunikasjon av usikkerhet som tilrettelegger for menneske-maskin-interaksjon. Videre skal oversikten synliggjøre relevante internasjonale standarder, som ISO, DNV og NIST, og deres anvendelse for å sikre at KI-løsninger overholder forskrifter og oppnår høy sikkerhet.

2.2 Bakgrunn

Petroleumsindustrien i Norge opererer faste og flytende installasjoner til havs, landanlegg og et omfattende rørledningssystem. Det er betydelige potensiale for forskjellige former for storulykker, eksempelvis relatert til branner, eksplosjoner, og oljelekkasjer. Av den grunn er det et stort fokus på sikkerhet både for mennesker og miljø.

De overordnede rammene for arbeidet med sikkerhet er gitt av Havtils regelverk for olje- og gassvirksomhet som inneholder fem forskrifter. Havtil gir for hver forskrift, tolkninger og veiledninger som beskriver hvordan regelverket kan anvendes, herunder henvisninger til bransjestandarder som kan benyttes for å oppfylle kravene. Det er aktørenes ansvar å følge regelverket som i stor grad er formulert som funksjonskrav, med fokus på behov og ønskede resultater, fremfor å foreskrive konkrete løsninger. Det betyr at aktørene i industrien i prinsippet har ganske stor frihet mht. til hvordan kravene skal møtes, men det at industrien har blitt enige om bruk av spesifikke standarder og veiledninger som er referert i Havtils regelverk har likevel bidratt til en høy grad av standardisering på tvers av aktører.

Et grunnleggende prinsipp som ligger til grunn for mange av kravene i regelverket er at svikt i en komponent, et system eller en enkelt feilhandling ikke skal føre til uakseptable konsekvenser. Det skal være uavhengige barrierer til stede som hinder at slike hendelser eskalerer og leder til ulykker. Relatert til dette har Havtil utgitt et notat knyttet til barrierestyring /7/ og et notat knyttet til risikostyring /8/ som gir ytterligere veiledning, utover veiledningene som ligger i selve regelverket.

KI blir nå innført i samfunnet i et relativt høyt tempo noe som gir utfordringer for arbeidet med sikkerhet i petroleumsindustrien. Det finnes så langt ingen standarder eller veiledninger som er utarbeidet spesifikt med tanke på sikker bruk av KI innenfor petroleumsindustrien. Eksisterende standarder og veiledninger tar hensyn til at forskjellige typer av programvare kan levere resultater som er feil eller ikke formålstjenlige, men er ikke blitt utarbeidet med KI i tankene. Dermed er det en mulighet for at måten man i dag håndterer programvarerelatert risiko på, ikke vil være effektiv for KI-baserte funksjoner.

I tillegg har Norge blitt EUs viktigste leverandør av gass, noe som betyr at evnen til å kontinuerlig levere gass har blitt enda viktigere enn før. Her ligger det en utfordringer i at Havtils regelverk fokuserer på sikkerhet og i liten grad på at systemene skal være tilgjengelige for produksjon. Det stilles ikke strenge krav til systemer som kun anses som kritiske for tilgjengeligheten av produksjon og leveranser, og terskelen for å ta i bruk KI for systemer som kun anses som tilgjengelighetskritiske vil dermed være betydelig lavere enn for sikkerhetskritiske systemer. Dette kan potensielt gi

positive effekter både når det gjelder tilgjengelighet og volum av leveranser, men risikoen for at KI gir negativ påvirkning må håndteres. Relatert til dette må aktører i petroleumsindustrien som opererer infrastruktur som er kritisk for gassforsyningen til EU forberede seg på å møte de nye reguleringskravene i EUs forordning om KI /50/.

Utover dette har det gjennom mange år vært en utvikling der graden av automatisering i industrien øker og der beslutninger i større grad tas basert på rapportering fra programvare som analyserer store datamengder. Dette betyr at det tradisjonelle skillet mellom operasjonsteknologi (OT) og informasjonsteknologi (IT) stadig blir mindre skarpt, og at systemer som tidligere kun ble ansett som rådgivende er blitt mer kritiske. Innføring av KI antas å forsterke denne trenden, ved at det kan bli vanskeligere å oppdage feil eller unøyaktige resultater fra programvare.

Innføring av KI vil også utfordre metodikk for teknologikvalifisering siden resultater fra KI-basert programvare i noen tilfeller ikke vil være deterministiske. Det betyr at dersom man gjennomfører samme test flere ganger, så vil man kunne få forskjellige resultater, noe som gir utfordringer knyttet til verifikasjon og validering.

I sum betyr dette at petroleumsindustrien har et stort behov for å utarbeide retningslinjer for hvordan KI-relatert risiko bør håndteres for forskjellige typer av aktiviteter, og denne kunnskapsoversikten er ment å gi et bidrag til dette arbeidet.

2.3 Metode

Det har blitt utført et omfattende litteratursøk i anerkjente databaser og fagfellevurderte tidsskrifter for å finne relevante artikler, rapporter, og bøker om KI og sikkerhetsaspekter i industrien. Søkeord som "kunstig intelligens", "storulykkesrisiko", og "sikkerhetskritiske systemer" ble brukt for å sikre dekkende resultater.

Videre har vi studert nasjonalt og internasjonale regelverk samt standarder og veiledninger som er relevante for bruk av KI i petroleumsindustrien. En viktig del av metoden var identifisering og evaluering av relevante standarder og rammeverk utarbeidet spesifikt for KI, men vi har også identifisert relevante standarder og veiledninger som ikke er utarbeidet med KI i tankene.

Vi vurderte også eksisterende praksiser for risikostyring og bruk av barrierer. Dette inkluderte vurderinger av hvordan dagens strategier for håndtering av programvarerelatert risiko kan tilpasses for implementering av nye KI-baserte løsninger.

KI i form av språkmodell har også blitt brukt i arbeidet, men alt i den endelige utgaven er gjennomgått og kvalitetssikret av mennesker.

2.4 Omfang og begrensninger

Siden denne studien handler om bruk av KI innenfor petroleumssektoren som storulykkes-industri, fokuserer denne rapporten på samme måte som Havtils regelverk på sikkerhet for menneske og miljø.

Dette er en vesentlig avgrensning, siden dette gjør at et hovedfokus i rapporten vil bli barrierefilosofi ved bruk av KI. En rapport som har et hovedfokus på å oppnå høyest mulig driftssikkerhet og verdiskapning med KI-løsninger ville hatt en annen struktur.

2.5 Definisjoner

Norsk begrep	Engelsk begrep	Definisjon
Barriereelement	Barrier element	Med barriereelement, menes tekniske, operasjonelle eller organisatoriske tiltak eller løsninger som inngår i realiseringen av en barrierefunksjon. Med tekniske barriereelementer menes utstyr og systemer som inngår i realiseringen av en barrierefunksjon. Med organisatoriske barriereelementer menes personell med definerte roller eller funksjoner og spesifikk kompetanse som inngår i realiseringen av en barrierefunksjon. Med operasjonelle barriereelementer menes de handlingene eller aktivitetene som personellet må utføre for å realisere en barrierefunksjon."
Forklarbarhet	Explainability	Egenskap ved et KI-system til å uttrykke viktige faktorer som påvirker KI-systemets resultater på en måte som mennesker kan forstå (ISO/IEC 22989, § 3.5.7).
Generativ KI	Generative AI	Refererer til modeller som brukes for å generere nye data. Disse modellene har som mål å lære den underliggende distribusjonen av dataene for å produsere nye eksempler som ligner på treningsdataene.
Informasjonsteknologi	Information technology	Bruken av datamaskiner og telekommunikasjonsutstyr for å lagre, hente, overføre og manipulere data. Det omfatter håndtering og behandling av data, inkludert maskinvare, programvare, nettverk og databaser som muliggjør informasjonsbehandling.
Kunstig intelligens	Artificial intelligence	Et systems evne til å utføre funksjoner som generelt er knyttet til menneskelig intelligens som resonnement og læringsevne (ISO/IEC 2382-28).
Livsløp	Life cycle	Evolusjon av et system, et produkt, en tjeneste, et prosjekt eller en annen menneskeskapt entitet, fra opprettelsen til avviklingen (ISO/IEC 22989, § 3.1.22).
Operasjonsteknologi	Operation technology	Maskinvare og programvare som brukes til å detektere eller forårsake endringer gjennom direkte overvåking og kontroll av industrielle anlegg.
Robusthet	Robustness	Et systems evne til å opprettholde ytelsesnivået under alle omstendigheter (ISO/IEC 22989, § 3.5.12).
Skjevhet	Bias	Systematisk avvik eller urettferdighet i resultatene fra en KI-modell, ofte som følge av dataene som brukes til trening eller designen av modellen.
Åpenhet	Transparency	Funksjonalitet som gir brukerne hensiktsmessig informasjon om KI systemets situasjonsforståelse og planlagte handlinger.

2.6 Forkortelser

Forkortelse	Definisjon
HCD	Menneskesentrert design / Human Centered Design
IEC	International Electrotechnical Commission
IKT	Informasjon- og kommunikasjonsteknologi
ISO	International Organization for Standardization
IT	Informasjonsteknologi
KI	Kunstig intelligens
LLM	Store språkmodeller / Large Language Models
ML	Maskinlæring
OT	Operasjonsteknologi
RRF	Risikoreduksjonsfaktor
V&V	Validering og verifikasjon

3 INTRODUKSJON TIL KUNSTIG INTELIGENS

Kunstig intelligens (KI) omfatter en rekke teknologier designet for å etterligne menneskelige kognitive funksjoner, med evnen til å lære, resonnere, løse problemer og ta beslutninger. For den norske offshoreindustrien, kan KI-applikasjoner bidra til fremskritt innen kostnadsreduksjon, operasjonell sikkerhet, effektivitet og beslutningsstøtte.

3.1 Hva er KI?

Det finnes mange definisjoner for hva kunstig intelligens er. I Tabell 3-1 er det listet noen definisjoner fra relevante standarder som viser at det er ganske stor spredning i hvordan det er definert.

Tabell 3-1 Definisjoner av KI.

Standard	Definisjon
ISO/IEC 2382:2015, ISO/IEC 29140:2021, ISO/TR 24291:2021	Branch of computer science devoted to developing data processing systems that perform functions normally associated with human intelligence, such as reasoning, learning, and self-improvement.
NS 11041:2024, ISO/IEC 2382-28:1995	Et systems evne til å utføre funksjoner som generelt er knyttet til menneskelig intelligens som resonnement og læringsevne.
DNV-RP-0671	Computer programs that generate output based on learnt transformations of data or other forms of automated reasoning.
Nasjonal strategi for kunstig intelligens	Kunstig intelligente systemer utfører handlinger, fysisk eller digitalt, basert på tolkning og behandling av strukturerte eller ustrukturerte data, i den hensikt å oppnå et gitt mål.
EU AI Act, EU-forordningen om kunstig intelligens	AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

Disse ulike definisjonene av kunstig intelligens (KI) har en del felles elementer, men også at de adresserer forskjellige aspekter av KI. Essensen og fellestrekkene i disse definisjonene er beskrevet nedenfor:

- **Menneskelig intelligens som referanse**: Flere definisjoner refererer til funksjoner som er assosiert med menneskelig intelligens, som resonnement, læring, og problemløsning. Dette er et gjennomgående tema som indikerer at KI handler om å etterligne menneskelige kognitive evner.
- **Autonomi og tilpasning**: Flere av definisjonene (spesielt ISO/TR 24291 og EU AI Act) fremhever KI-systemers evne til å operere autonomt og tilpasse seg nye situasjoner basert på læring fra data. Dette betyr at KI-systemer ikke bare følger forhåndsdefinerte regler, men kan også tilpasse seg basert på erfaringer.
- **Databehandling og læring**: KI-systemer er knyttet til behandling av data, enten strukturerte eller ustrukturerte, og bruken av denne dataen for å generere innsikt, anbefalinger, eller handlinger. Læring / trening fra data er sentralt i flere av definisjonene.
- **Målrettethet**: Definisjonene fremhever også at KI har et formål eller mål som systemet forsøker å oppnå, enten dette er eksplisitt spesifisert eller implisitt.

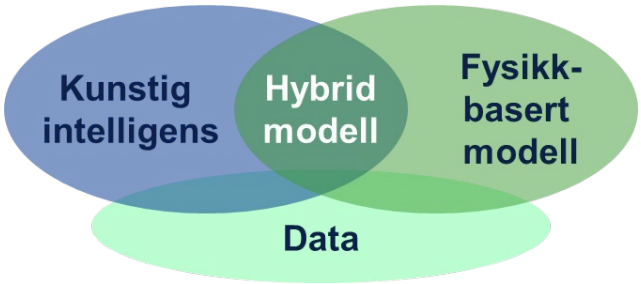
Flere av definisjonene, spesielt de fra Nasjonal strategi for kunstig intelligens og EU AI Act, er brede nok til også å inkludere digitale systemer som ikke tradisjonelt ville blitt sett på som KI, som for eksempel simuleringsmodeller. Dette skyldes fokuset på systemers evne til å tolke data og påvirke sine omgivelser, noe som kan inkludere en rekke andre digitale teknologier.

Denne kunnskapsoversikten vurderer i prinsippet alt som faller inn under en av definisjonene ovenfor, men vi fokuserer spesielt på de typene som er identifisert i kapittel 3.2 nedenfor, siden disse anses som mest relevant for systemer med storulykke potensial.

3.2 Typer av kunstig Intelligens

Noen eksempler på typer av kunstig intelligens relevant for denne kunnskapsoversikten:

- **Maskinlæring (ML)** gjør at systemer kan lære fra data, identifisere mønstre og ta beslutninger med minimal menneskelig inngripen. Relevante eksempler er ML-applikasjoner for prediktivt vedlikehold, hvor algoritmer forutser spesifikke typer av utstyrsfeil før de oppstår, og ML-applikasjoner for anomalideteksjon i store menger operasjonsdata,
- **Generativ KI** refererer til modeller som brukes for å generere nye data. Disse modellene har som mål å lære den underliggende distribusjonen av dataene for å produsere nye eksempler som ligner på treningsdataene.
- **Diskriminativ KI-modeller (discriminative AI)** er en type kunstig intelligens som brukes for prediksjon og inferens basert på eksisterende data. De genererer ikke nye data, men fokuserer på å identifisere mønstre og ta beslutninger basert på disse mønstrene.
- **Store språkmodeller (LLM)** er en type generativ KI som er trent på enorme mengder tekstdata for å kunne forstå, generere og behandle naturlig språk på en menneskelig måte. Modeller som GPT (Generative Pretrained Transformer), bruker dype nevrale nettverk for å fange opp mønstre, grammatiske regler, betydninger og kontekst i språket.
- **Regel-baserte modeller** omfatter systemer som følger forhåndsdefinerte regler eller logikk for å ta beslutninger, potensielt i kombinasjon med maskinlæring. For eksempel kan regelbaserte systemer overvåke operasjonelle parametere i sanntid og sende ut varsler eller iverksette korrektive tiltak når avvik oppdages.
- **Hybride KI-modeller** integrerer datadrevne tilnærminger med grunnleggende fysiske prinsipper (deterministiske eller statistiske) for å modellere og forutsi systematferd. Eksempler inkluderer optimalisering av boreoperasjoner basert på geologiske modeller og sanntidsdata, og dynamiske simuleringen av hvordan offshore-konstruksjoner vil respondere på påvirkning fra det operative miljøet.



Figur 3-1 Forhold mellom KI og fysikk-baserte modeller.

Merk at anvendelsene nevnt ovenfor har et fokus på KI brukt i direkte støtte til operasjoner. Det finnes en rekke andre anvendelser som vil være relevante for petroleumsindustrien. For eksempel må man forvente at KI vil bli brukt til å automatisk generere kildekode både for bruk i kontrollsystemer, og for bruk i rådgivende systemer.

De ulike typene KI tilbyr verktøysett som har potensiale til å forbedre sikkerhet og operasjonell effektivitet i petroleumsindustrien. Fremskritt innen ML og LLM-er åpner for nye muligheter i dataanalyse og beslutningsstøtte, mens regelbaserte og fysikk-baserte KI-applikasjoner støtter beslutningstaking basert på etablert kunnskap og prinsipper.

3.3 Norge og KI

KI forventes å spille en stadig viktigere rolle i Norge, og regjeringen utarbeidet i 2020 en nasjonal KI-strategi /49/, som tar sikte på å fremme innovasjon samt bruk av KI-systemer på en ansvarlig og sikker måte. Strategien fokuserer på blant annet transparent bruk, datasikkerhet, og etisk utvikling av KI-teknologier.

Regjeringen i Norge har satt av én milliard kroner som skal gå til forskning på kunstig intelligens og digital teknologi. Forskningsmilliarden vil bidra til større innsikt om hvilke konsekvenser teknologiutviklingen har for samfunnet. Det vil også gi mer kunnskap om nye digitale teknologier og muligheter for innovasjon i næringslivet og offentlig sektor. Satsingen blir finansiert innenfor Kunnskapsdepartementets ramme. Forskningsmilliarden vil ha tre hovedspor

- Samfunnsrelatert forskning slik som lovregulering, demografi, personvern osv.
- Forskning på digitale teknologier, slik som neste generasjon IKT, digital sikkerhet, datakvalitet osv.
- Forskning på hvordan digitale teknologier kan brukes på innovasjon i privat og offentlig næring.

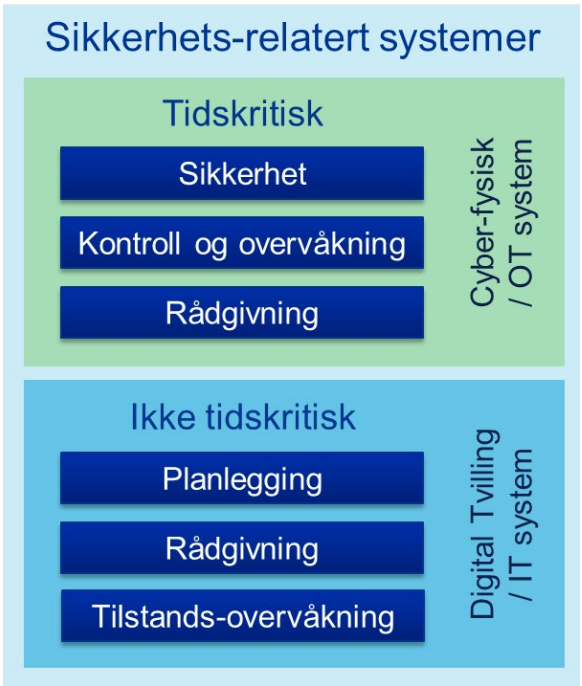
Det finnes allerede etablerte senteret for KI-forskning, og flere er planlagt under forsknings-milliarden. Eksempler på allerede etablerte sentere er NorwAI, NAIL, CAIR, dScience, BigInsight, NORA og Integreat.

4 RISIKO OG SÅRBARHETER KNYTTET TIL UTVIKLING OG BRUK AV KI

Dette kapittelet illustrerer at det er mange typer av systemer som vil kunne ha en påvirkning på sikkerheten, selv om de ikke er definert som rene sikkerhetssystemer. Dette er viktig siden de strenge kravene som stilles til rene sikkerhetssystemer gjør at terskelen for å innføre bruk av KI i slike systemer vil være svært høy, mens den vil være betydelig lavere i andre typer av sikkerhetsrelaterte systemer. Videre ser dette kapittelet på bruk av KI i et systemperspektiv der systemet består av menneske, teknologi og organisasjon, det indentifiserer typiske applikasjoner hvor KI forventes tatt i bruk på relativt kort sikt, og identifiserer et antall risikoer knyttet til disse.

4.1 Sikkerhetsrelaterte systemer

I dette dokumentet brukes begrepet sikkerhetsrelaterte systemer som er illustrert i Figur 4-1. Det refereres til programmerbare logiske systemer og ikke til mekaniske systemer slik som en trykksikringsventil. Selv om sikkerhetssystemene primært har ansvaret for sikkerhetsfunksjonene så vil de andre systemene også kunne påvirket potensialet for en storulykke hvis de ikke fungerer slik de er tiltenkt. Figuren skiller mellom tidskritiske og ikke-tidskritiske systemer, hvor tidskritiske systemer refererer til prosesser eller oppgaver som krever en rask handling eller gjennomføring innenfor begrenset tidsramme, mens ikke-tidskritisk systemer ikke krever en rask handling. Tidskritiske system vil typisk implementeres som en del av operasjonsteknologi (OT) systemet, som inkluderer prosess-kontroll, nød-avstengning, og sikkerhetsfunksjoner.



Figur 4-1 Sikkerhetsrelaterte systemer.

- **Sikkerhets-systemer** – et programmerbart system som integrerer deteksjonsmekanismer, logiske enheter og aktuatorer med tanke på å oppdage farlige forhold og bringe en prosess eller system til en sikker tilstand. Disse systemene er designet for å redusere risikoen for farlige hendelser ved at det automatisk eller manuelt blir iverksatt forhåndsdefinerte aksjoner, slik som nød-avstengning eller trykkavlastning. Relevante standarder her inkluderer NORSOK S-001, IEC 61508, IEC 61511, Offshore Norges retningslinje 070, ISO 13702, ISO 13849, og disse setter strenge krav til hvordan slike systemer blir designet, verifisert og driftet.

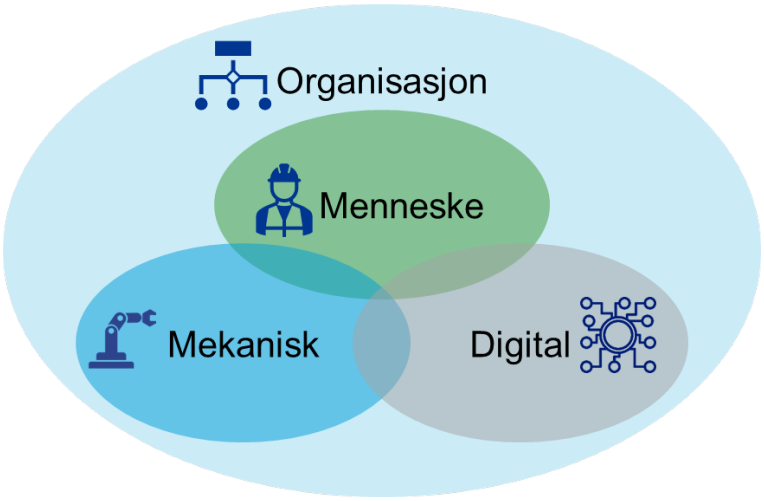
- **Kontroll-systemer** – et automatisert system som styrer, overvåker og regulerer prosesser og operasjoner. Eksempler er Prosess-kontrollsystemer, boresystemer og andre automasjonssystemer. Relevante standarder inkluderer NORSOK P-002, NORSOK I-001, NORSOK I-002, ISO 10418.
- **Rådgivnings-systemer** – system som gir beslutningsstøtte ved å analysere data for å støtte operatørens situasjonsforståelse og foreslå videre aksjoner. Rådgivningssystemer kan brukes til å optimalisere operasjoner, forutsi vedlikeholdsbehov, og forbedre sikkerhet og effektivitet. Disse systemene kan kombinere ekspertkunnskap med dataanalyse for å gi operatører og ledere bedre innsikt i en situasjon og kommer med anbefalinger om tiltak. Rådgivningssystemer kan benyttes både i tidskritiske og ikke-tidskritiske systemer, men her er det få etablerte standarder.
- **Planleggings-systemer** – system som brukes til å planlegge operasjoner og vedlikehold som har potensiale for ulykker, slik som boreoperasjoner, løfteoperasjoner, vedlikehold på prosessanlegget, etc. Dette innebærer å koordinere alle aktiviteter for å unngå uforutsette problemer, sikre at alle ressurser brukes effektivt, samtidig som sikkerheten ivaretas. Her er det få etablerte standarder.
- **Tilstandsovervåkning-systemer** – et system som kontinuerlig samler inn, analyserer og rapporterer data fra ulike prosesser og utstyr. Disse systemene brukes til å oppdage avvik, feil, analysere trender, osv. for å forutsi når utstyr vil feile for å planlegge vedlikehold eller utskifting. Dette har til hensikt å tidlig identifisere mulige problemer og sikrer kontinuerlig drift og sikkerhet. Relevante standarder inkluderer DNV-RP-A204 og ISO 13374.

4.2 KI i et systemperspektiv

I et industrielt perspektiv så vil KI ofte bli integrert inn i et system som består av et mekanisk system, et digitalt system hvor KI kan inngå, og mennesker som opererer systemet. I tillegg så vil organisasjonene som utvikler og vedlikeholder systemet, inklusiv styring-systemet, påvirke hvor sikkert og pålitelig det er i drift. Innen petroleumsnæringen blir dette omtalt under begrepet MTO (Menneske, Teknologi og Organisasjon):

- **Menneske**: Dette fokuserer på menneskelige faktorer som kompetanse, erfaring, arbeidsmiljø og hvordan menneskelige feil eller suksesser kan påvirke sikkerheten og effektiviteten i organisasjonen. Det handler også om hvordan mennesker samhandler med teknologi og organisasjonsstrukturer.
- **Teknologi**: Teknologien refererer til utstyret, maskineriet, programvaren og systemene. Teknologiperspektivet ser på hvordan teknologien brukes, hvordan den påvirker arbeidsprosesser, hvordan den kan påvirke situasjonsforståelsen til operatøren, og hvordan den kan optimaliseres for sikkerhet og effektivitet.
- **Organisasjon**: Dette omhandler organisasjonsstrukturer, prosedyrer, kultur og ledelse. Organisatoriske faktorer kan ha stor innvirkning på både menneskers prestasjoner og hvordan teknologi blir brukt. Effektiv kommunikasjon, klare ansvarsområder og gode rutiner er viktige aspekter her.

MTO-perspektivet er spesielt viktig for å utvikle et helhetlige system som er sikkert og robust. Når det gjelder teknologidelen så er det fornuftig å dele den inn i det mekaniske og digitale systemet da disse typisk er utviklet av forskjellige miljøer og hvor samhandlingen mellom dem er kritisk å forstå og analysere, se Figur 4-2. Samspillet mellom disse elementene, hvor KI også integreres, gjør dette til et komplekst system.



Figur 4-2 Samspillet mellom mennesker, teknologi og organisasjonen.

KI kan i prinsippet være en del av alle sikkerhetsrelaterte systemer (se Figur 4-1) men tiltakene for å kvalifisere KI-løsningen må være proporsjonal med kritikaliteten til systemet, og av den grunn ser DNV det som lite sannsynlig at KI på kort sikt vil bli innført i det som innenfor petroleumsindustrien klassifiseres som sikkerhetssystemer.

Det vil ikke være tilstrekkelig å kvalifisere KI-modellen som en uavhengig modul, den må kvalifiseres ut ifra kravene som systemet som helhet krever, og i noen tilfeller kan kvalifisering av KI også drive endringer i systemet. Dette inkluderer også krav til å formidle til brukerne av systemet om informasjonen som produseres har tilstrekkelig kvalitet og troverdighet.

4.3 Bruk av barrierer

Barrierer er tiltak som skal forhindre at farer eller ulykker skjer, eller redusere konsekvensene hvis noe likevel skjer. Barriereelementer kan være tekniske (som utstyr og digital systemer), organisatoriske (styringsmodeller og prosedyrer) eller operasjonelle (menneskelige handlinger).

En «bow-tie» risikomodell brukes ofte for å visualisere risikoen for større hendelser, både fysiske og digitale. Den identifiserer systematisk trusler og barrierer (venstre side) mot en uønsket hendelse (midten), og barrierer som kan forhindre eskalering av hendelsen og dermed redusere konsekvensene (høyre side). Figur 4-3 viser en forenklet utgave av en slik «Bow-tie» modell for KI.

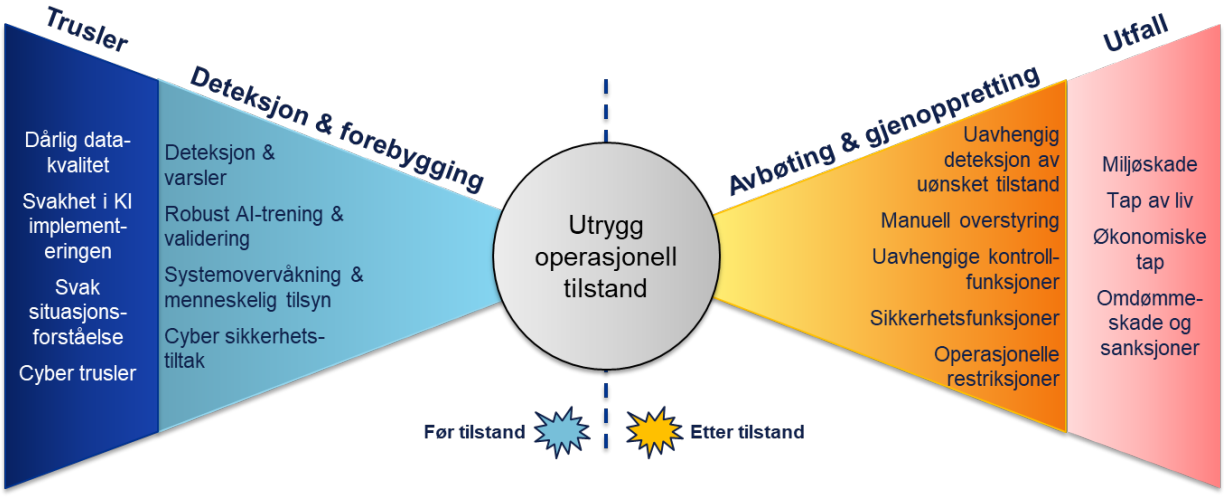
Havtil setter krav til barrierer og barrierestyring, som handler om å sikre at alle barrierer fungerer som de skal gjennom hele levetiden til en installasjon eller operasjon. Barrierestyring (basert på sikkerhetsbarrierer og barriere-styrings notat) handler om:

- Indentifisering av trusler og svekkelser.
- Definering av nødvendige barrierer for å håndtere truslene.
- Vedlikehold, testing og validering av barrierer for å sikre at de fungerer som de skal.
- Oppfølging og overvåkning av barrierenes tilstand, både teknisk og organisatorisk.

Ytelsespåvirkende faktorer for barrierer refererer til forhold som kan påvirke hvor godt en barriere fungerer, altså dens evne til å utføre sin tiltenkte funksjon under normale eller kritiske forhold. I barrierehåndtering handler det om å forstå og kontrollere disse faktorene slik at barrierens ytelse ikke svekkes. Noen sentrale ytelsespåvirkende faktorer for barrierer er:

- Tekniske faktorer som slitasje, vedlikehold, aldring av utstyr, sensor feil, etc.
- Menneskelig faktorer som kunnskap og ferdigheter, trøtthet og stress, tid tilgjengelig til handling, situasjonsforståelse, og kommunikasjon.
- Organisatoriske faktorer som ledelse, kultur, prosedyrer og retningslinjer, og resurser.
- Miljømessige faktorer som miljølaster, korrosivt miljø, temperatur, etc. som kan påvirke ytelsen.
- Interaksjon mellom barrierer som for eksempel en feil i en teknisk barriere kan kreve at en annen operasjonell barriere reagerer svært raskt, som for eksempel en handling fra en operatør..

KI i industrielle systemer kan både være en mulig trussel og / eller fungere som en barriere mot en trussel. Uansett så er det essensielt å vurdere KI applikasjonen ut ifra hvordan den påvirker systemet og hvordan det passer inn i ett barriere-perspektiv.



Figur 4-3 Bow-tie modell for KI-systemer.

Innholdet i de forskjellige elementene i figuren er som følger:

Trusler (årsaker som kan lede til uønsket operasjonell tilstand)

Eksempler på årsaker som kan lede til uønsket operasjonell tilstand er som følger:

- Utilstrekkelig KI-trening eller skjevhet som fører til at KI feiltolker situasjoner eller produserer usikre beslutninger.
- Dårlig datakvalitet som fører til at KI-systemet tar feil beslutninger.
- Modellforringelse er noe som oppstår når de forholdene en KI-algoritme ble trent under, gradvis endrer seg over tid, eller når uventede variabler introduseres i det operasjonelle miljøet.
- Overtilpassing til treningsdata, hvor en KI-modell tilpasser seg for godt til treningsdata som inkluderer støy og detaljer som ikke er relevante. Dette gjør at modellen presterer godt basert på treningsdata, men presterer dårlig på nye data.
- Feil som blir introdusert i interaksjonen mellom KI og menneske.

- Feil i implementering av applikasjonen som inneholder KI-algoritmen, og/eller feil i implementering av annen programvare i den relevante teknologistacken, eksempelvis i operativsystemet.
- Hardware feil i maskinvaren KI-algoritmen kjører på. Det kan skyldes såkalte tilfeldige hardware feil, eller negativ påvirkning fra omgivelsene, eksempelvis temperaturøkning eller elektromagnetisk stråling.
- Villede handlinger som påvirker KI-algoritmen, eksempelvis i form av hacking, datavirus, jamning av kommunikasjon eller sabotasje.

Deteksjon og forebygging (barrierer for å forhindre at uønsket tilstand oppstår)

Eksempler på forbyggende barrierer er som følger:

- Testing og validering av KI-modeller regelmessig, for å sikre at de fungerer riktig i virkelige situasjoner.
- Sanntidsovervåking og varsler: Kontinuerlig overvåking av systemytelse, med sanntidsvarsler for unormale hendelser.
- Kontinuerlig overvåking av innkommende data for å sjekke datakvaliteten.
- Sikkerhetsbeskyttelse: Brannmurer, kryptering og flerlags sikkerhetsprotokoller for å forhindre uautorisert tilgang til KI-systemet.
- Validering av KI-generert informasjon ved hjelp av andre verktøy som generer tilsvarende informasjon på en annen måte. Eksempelvis kan det være relevant å bruke forskjellige typer simuleringsverktøy, med og uten KI.
- Bruken av KI begrenses til å være rådgivende, slik at mennesker er involvert i kritisk beslutningstaking.
- Usikkerhetskvantifisering som innebærer implementering av metoder som kvantifiserer usikkerhet i KI-modeller. Det gjør at prediksjoner kan tilordnes et konfidensnivå.

Utrygg operasjonell tilstand

Dette er det sentrale elementet i modellen (sirkelen i midten av figuren). Den representerer en situasjon der det KI-baserte systemet har levert data eller en kommando som gjør den operasjonelle situasjonen utrygg. For eksempel kan den operasjonelle situasjonen representer en sjelden hendelse som KI-systemet ikke er trent for.

Avbøting og gjenoppretting (barrierer som hindrer eskalering til farlig hendelse)

Eksempler på barrierer er som følger:

- Manuell overstyring tillater menneskelige operatører å overstyre KI-beslutninger i kritiske øyeblikk.
- Bruk av kontrollfunksjoner som er uavhengige av applikasjonen som inneholder KI, og som er kapable til holde prosessen som blir kontrollert innenfor sikker tilstand. Slike funksjoner vil bli aktivert basert på manuell eller automatisert deteksjon av den uønskede tilstanden.
- Bruk av sikkerhetsfunksjoner som stenger ned prosessen som blir kontrollert. Slike funksjoner vil bli aktivert basert på manuell eller automatisert deteksjon av den uønskede tilstanden.
- Operasjonelle restriksjoner som reduserer konsekvens, selv om eskalering skulle skje. Et eksempel på dette er at mennesker fysisk utestenges fra området der en operasjon foregår.
- Andre former for tiltak som reduser konsekvens, selv om eskalering skulle skje, eksempelvis brannbekjempelse.

Merk at manuell overstyring ved hjelp av programvare som kjører på samme maskinvare som KI-algoritmen, ikke vil være en uavhengig barriere, uansett hva som er årsak til den uønskede tilstanden.

I tillegg vil alle de tre førstnevnte typene av barrierer være avhengige av å kunne detektere den uønskede tilstanden uavhengig av systemet som inneholder KI, dersom barrierene skal være effektive uansett årsaken til problemet. Dette må også skje så fort at barrieren blir aktivert tidsnok til å ha en effekt.

Eksempelvis er det ikke gitt at man får en alarm fra systemet som inneholder KI dersom årsaken til den uønskede hendelsen er at KI-algoritmen ikke er trent for et operasjonelt scenario som inntreffer sjelden, eller dersom årsaken er en implementeringsfeil. Dersom man ikke kan detektere den uønskede tilstanden på andre måter, vil eskaleringen i slike tilfeller kunne være et faktum.

Dette deteksjonsproblemet forventes i mange tilfeller å være hovedutfordringen når det gjelder sikker bruk av KI, særlig når man er avhengig av menneskelig deteksjon. Problemstillingen er ikke ny, og vil ofte være til stede i forbindelse med automatisering av prosesser, uavhengig av om KI blir benyttet eller ikke, men KI kan i noen tilfeller forsterke problematikken. Se også kapittel 5.2 og 5.7.

Utfall (potensielle konsekvenser dersom den uønskede hendelsen eskalerer)

Verst tenkelige utfall vil variere sterkt avhengig av type KI applikasjon, men man kan tenke seg farlige brønnhendelser, offshore kraner som beveger seg på en farlig måte, uønskede utslipp osv.

4.4 Eksempler på applikasjoner ansett som relevante i denne studien

Bruken av KI i energisektoren, spesielt innen offshoreindustrien, er i startfasen, men står foran en stor forventet vekst. Noen eksempler på relevante applikasjoner er gitt nedenfor:

- **Applikasjoner som gir beslutningsstøtte:** I den rådgivende enden av spekteret er KI-systemer designet for å støtte menneskelig beslutningstaking, snarere enn å erstatte den. Disse applikasjonene fokuserer hovedsakelig på å tilby innsikt, anbefalinger og forbedrede analyser som hjelper ingeniører og operatører i å ta informerte beslutninger. Eksempler inkluderer:
    - **Prediktivt vedlikehold:** Bruk av KI for å analysere data fra sensorer og maskineri for å forutsi utstyrssvikt før de oppstår, slik at vedlikehold kan utføres tidsnok og uforutsett nedetid reduseres /35/ /36/.
    - **Planlegging:** Avansert KI-analyse brukes for å forutsi vedlikeholdsbehov, optimalisere ressursallokering, og planlegge forebyggende tiltak.
    - **Sikkerhetsovervåking:** Implementering av KI-drevne overvåkningssystemer for å overvåke sikkerhets-samsvar og identifisere potensielle farer i sanntid.
    - **Energioptimering:** Bruk av KI-algoritmer for å analysere mønstre i energiforbruk og operasjonelle data, og anbefale justeringer for å optimalisere energibruk og redusere kostnader /38/ /39/.
- **Applikasjoner som benyttes i forbindelse med kontroll og overvåking:** Her begynner KI-applikasjoner å ta mer direkte kontroll over visse prosesser, samtidig som de fortsatt opererer under menneskelig tilsyn. Disse applikasjonene er kjennetegnet ved deres evne til å utføre spesifikke oppgaver autonomt basert på forhåndsdefinerte regler kombinerte med maskinlæring. Menneskelig intervensjon skjer kun når det oppstår uønskede tilstander og i forbindelse med strategiske beslutninger. Eksempler inkluderer:
    - **Automatiserte boreoperasjoner:** KI-systemer som kontrollerer boreutstyr, justerer parametere i sanntid for optimal ytelse, underlagt overordnet kontroll av menneskelige operatører /23/ /26/ /28/ /29/ /30/.

- **Intelligente kontrollsystemer:** KI-baserte kontrollsystemer som håndterer driften av ventiler, pumper, og andre kritiske infrastrukturkomponenter, forbedrer effektiviteten og reduserer risikoen for menneskelige feil /22/.

- **Autonome applikasjoner:** På den ytterste enden av spekteret er fullt autonome KI-applikasjoner, som opererer uavhengig av menneskelig intervensjon. Selv om dette nivået av autonomi ennå ikke er utbredt i offshoreindustrien, representerer det et langsiktig mål for områder hvor menneskelig nærvær er spesielt risikabelt, eller hvor KI kan yte betydelig bedre enn menneskelige kapasiteter. Eksempler på områder hvor KI kan forventes å bli introdusert er:

  - **Autonome undervannsfarkoster (AUV-er) for inspeksjon:** AUV forventes utstyrt med KI for å autonomt navigere og inspisere tilstanden til undervannsinfrastruktur, uten direkte menneskelig kontroll /40/ /41/. Skulle en slik farkost kollidere er skadepotensiale i mange tilfeller lavt. Derfor forventes det at terskelen for å innføre KI i den autonome navigasjonen av slike farkoster også vil være tilsvarende lav.

4.5 Eksempler på risiko knyttet til bruk av KI

Typiske årsaker til at KI produserer uønskede resultater kan være: forringelse av modeller, skjevhet i data, skjevhet i informasjonsmodeller, for dårlig datakvalitet, og mangel på gode data for trening av algoritmene. Mangel på data fra sjeldne hendelser kan være et eksempel på det siste. Se kapittel 5.6 for mer detaljer om typiske årsaker.

I hvilken grad informasjon produsert ved hjelp av KI kan lede til storulykke, er avhengig av hva informasjonen brukes til, i hvilken grad det er mulig å detektere problemet på en uavhengig måte, og om det er tid nok til både å detektere og iverksette tiltak. Se også 4.3 når det gjelder bruk av barrierer.

Noen eksempler der bruk av KI kan introdusere risiko er nevnt nedenfor.

- KI forventes tatt i bruk i forbindelse med prediktivt vedlikehold og planlegging av vedlikehold. Dersom KI blir brukt til å bestemme når vedlikehold skal utføres, og dette kan skje sjeldnere enn det som har vært normen så langt, ligger det en risiko for at vedlikehold som burde ha vært utført ikke blir utført, med påfølgende økt ulykkesrisiko.

- Bruk av rådgivende applikasjoner som baserer seg på kontinuerlig innsamling og analyse av data fra forskjellige typer systemer kan også bli brukt som et argument å øke testintervaller, noe som kan gi en risiko for at farlige feilsituasjoner blir oppdaget senere enn det som ville vært tilfelle ved hyppigere testintervaller.

- KI forventes tatt i bruk både i kontrollsystemer og i applikasjoner som gir råd om hva som bør gjøres i spesifikke operasjoner. Eksempler er applikasjoner og systemer som brukes i forbindelse med løfteoperasjoner og boreoperasjoner. Her ligger det en risiko knyttet til at informasjon produsert av KI kan være feil eller ikke-hensiktsmessig for spesifikke operasjonelle scenarioer. Dette kan være vanskelig for en operatør å detektere, noe som potensielt kan forårsake problemer i operasjon, samt at relevante nød-funksjoner ikke blir aktivert.

- KI forventes tatt i bruk i forbindelse med kravspesifisering, design, implementasjon, verifikasjon og validering av systemer. Dette vil være relevant både for systemer som selv benytter KI i sine algoritmer, og systemer som ikke gjør det. Det forventes at dette i mange tilfeller vil kunne øke kvaliteten på prosessene, for eksempel kan bruk av KI muliggjøre mer omfattende testing av systemer. Det ligger imidlertid en fare i at leverandørene kan komme til å stole for mye på verktøyene, slik at verifikasjon og valideringsprosesser som involverer mennesker blir for svake.

- KI forventes også tatt i bruk til å produsere mange forskjellige typer av dokumentasjon, eksempelvis operasjonelle prosedyrer, også her ligger det en fare i at verifikasjon og valideringsprosesser som involverer mennesker kan bli for svake.

5 METODER FOR Å SIKRE ROBUSTE LØSNINGER

5.1 Teknologikvalifisering

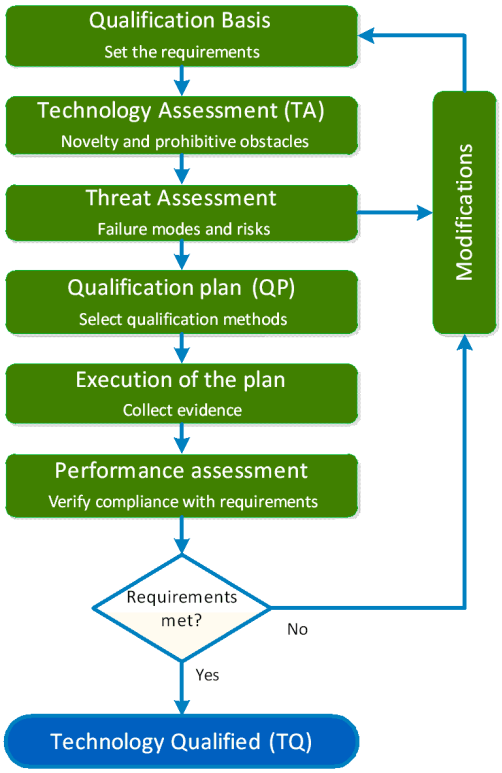
Innretningsforskriften Paragraf § 9, «Kvalifisering og bruk av ny teknologi og nye metoder», sier at:

> *Der petroleumsvirksomheten medfører bruk av ny teknologi eller nye metoder, skal det utarbeides kriterier for utvikling, testing og bruk slik at kravene til helse, miljø og sikkerhet blir oppfylt.*

I veiledning til denne paragraf 9 i innretningsforskriften står følgende:

> *For å oppfylle kravet til metode for kvalifisering av ny teknologi kan DNV-RP-A203 og Oil & Gas UK Guidelines on Qualification of Materials for the Abandonment of Wells, issue 2, brukes.*

For kvalifisering av KI vil det være DNV-RP-A203 som er mest relevant, og kvalifiseringsprosessen den anbefaler er vist i Figur 5-1.

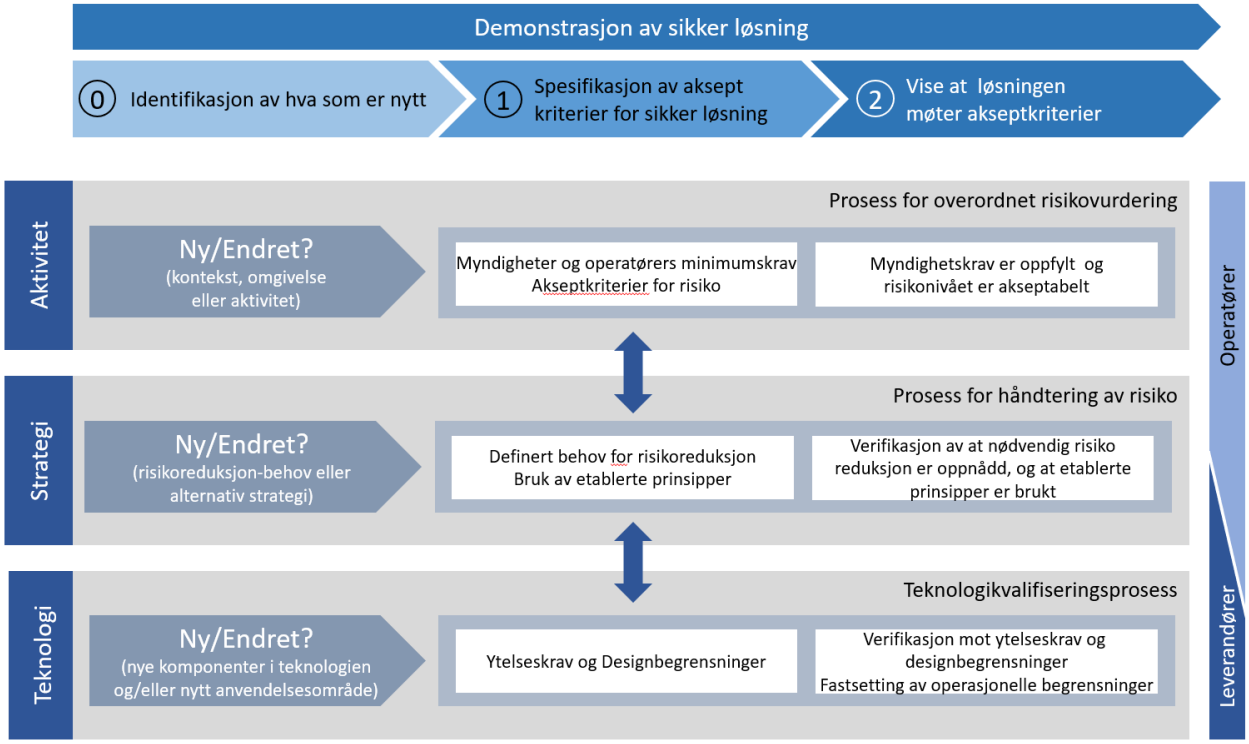


Figur 5-1 Overordnet TQ prosess fra DNV-RP-A203.

Prosessen er relevant for alle typer teknologi, men for et KI relevant prosjekt vil det av to årsaker ikke være tilstrekkelig å bruke denne veiledningen alene. Den viktigste er at DNV-RP-A203 er orientert mot kvalifisering av fysiske komponenter, og dermed har begrenset veiledning når det gjelder kvalifisering av programvarebaserte funksjoner. Det vil derfor være relevant å støtte seg på andre standarder og veiledninger identifisert i denne kunnskapsoversikten, se oversikt i kapittel 6.3.

Den andre årsaken til at innføring av KI vil kunne medføre endringer i hvordan man utfører en aktivitet og/eller utløse behov for en strategiendring, for eksempel når det gjelder hvilke uavhengige barrierer som er nødvendig å ha på plass. En demonstrasjon av sikkerhet kan derfor kreve endringer utover det å kvalifisere at den nye teknologien har den nødvendige ytelsen isolert sett.

Eksempelvis kan man tenke seg at innføring av en KI in en kontrollfunksjon eller i en rådgivende funksjon gir behov for en ny barrierestrategi der man går over fra menneskelig aktivering av uavhengig sikkerhetsfunksjon til fullautomatisert aktivering. Det betyr at kvalifisering av en KI-løsning potensielt kan utløse et behov for et mer avansert sikkerhetssystem med forskjellige typer sensorer og en mer avansert logikk enn den som er nødvendig når manuell aktivering benyttes.



Figur 5-2 Innføring av ny teknologi kan medføre endringer på flere nivåer.

Figur 5-2 illustrerer at det å demonstrere sikkerhet i forbindelse med innføring av ny teknologi kan medføre behov for endringer som går utover selve teknologikvalifiseringen. Den er hentet fra Safety 4.0 som var et Joint Industry Project (JIP) som så på hvordan man kan innføre ny undervannsteknologi på en sikker måte. I dette prosjektet deltok fire oljeselskaper, tre leverandører, universitetene i Trondheim og Stavanger samt DNV.

Rapportene fra Safety 4.0 er ikke laget med KI i tankene, men har i forhold til DNV-RP-A203 større fokus på systemperspektivet beskrevet i kapittel 4.2, samt mer veiledning når det gjelder håndtering av programvarerelatert risiko. Strategiene som i dag brukes til dette formålet er også aktuelle når det gjelder sikker bruk av KI, se kapittel 5.2 nedenfor for en oversikt over slike strategier. Resultatene fra Safety 4.0 prosjektet er også brukt i arbeidet med å utarbeide DNV-RP-0671 Assurance of AI-enabled systems /19/.

5.2 Håndtering av programvarerelatert risiko i eksisterende systemer

Svakheter som kan føre til at uønskede tilstander opptrer i programvare, kan introduseres på mange forskjellige punkter i livssyklusen til et system. Dette gjelder uavhengig av om KI blir benyttet eller ikke.

Svakheter kan oppstå allerede når man spesifiserer krav til og designer systemer. Dette kan skje fordi man overser avhengigheter mellom de tekniske, operasjonelle, menneskelige og organisatoriske komponenter, utarbeider spesifikasjoner som er basert på utilstrekkelig forståelse av fysiske prosesser og operasjonelt miljø, eller ikke tar hensyn til mulig uønsket innputt fra andre systemer.

Svakheter kan også introduseres på programvaredesign og implementeringsnivå, for eksempel gjennom uforutsette avhengigheter i den interne dataflyten, eller usikker bruk av programmeringsspråket. Programvaren kan også være for lite robust mot degradering i maskinvareplattformen der den kjøres. Dermed kan uønskede tilstander som er forårsaket av forringet, men fortsatt operativ maskinvare, fremstå som programvarerelaterte.

Svakhetene i programvaren vil enten være til stede i et system fra dag én i operasjon, eller bli introdusert gjennom modifikasjoner. De vil typisk være skjult inntil spesifikke omstendigheter inntreffer, og det finnes i dag ingen omforent metode som gjør at man kan beregne sannsynligheten for at:

- Alvorlige svakheter er til stede i programvaren, og i så fall antall slike svakheter.
- Verdier, kombinasjoner og/eller sekvenser av inngangsdata, som er nødvendige for å avdekke disse svakhetene faktisk forekommer under operasjon.

Videre, når man ser på settet med eksterne grensesnitt som er relevante for en programvarefunksjon, så kan antallet mulige kombinasjoner og sekvenser av inngangsdata ofte være ekstremt høyt. Dette betyr at uansett hvilken type teknologistakk som brukes, så er det vanligvis ikke mulig å teste programvaren uttømmende.

Testproblemet forsterkes vanligvis ytterligere av den interne utformingen av programvaren. Et eksempel er at antallet mulige veier gjennom applikasjonsprogramvaren ofte vil være praktisk talt uendelig på grunn av forskjellige former for tilbakekobling, eksempelvis i tilstandsmaskiner. Et annet eksempel er at kompleksiteten i dataflyten innenfor en programvare kan skjule avhengigheter mellom ulike deler av en programvare, noe som er en av grunnene til at uønskede bivirkninger av modifikasjoner er ganske vanlig.

Det at uttømmende testing ikke er mulig, gjør at testing ikke kan demonstrere fravær av svakheter i programvare. Testing er en svært viktig måte å redusere antall svakheter, men det faktum at alle tester utviklet for en sikkerhetskritisk programvare har bestått, er ikke alene tilstrekkelig bevis for sikkerhet. Dette gjør det generelt svært utfordrende å demonstrere at man har kontroll på risiko knyttet til programvare dersom det ikke finnes uavhengige barrierer som kan forhindre et problem fra å eskalere til en farlig hendelse.

Det store mulighetsrommet gjør også at det ofte er vanskelig å vite om kravspesifikasjon for en programvare er komplett og egnet til formålet, fordi man ikke klarer å overse hva slags scenarioer programvaren kan bli utsatt for i driftsfasen. Dette betyr også at det kan være usikkert om matematiske algoritmer som skal implementeres i programvaren er egnet til formålet under alle omstendigheter. Når det gjelder sistnevnte, kan innføring av KI introdusere ytterligere usikkerhet.

Driftserfaring blir ofte brukt som et argument for at programvare er trygg å bruke. Dette er fordi antall svakheter typisk vil bli redusert dersom en programvare har vært i drift over en lengre periode der endringer kun har vært drevet av feilretting, og det ikke har vært andre typer modifikasjoner. Det vil likevel være vanskelig å vurdere hvor mye risikoreduksjon dette gir, siden det i fremtiden kan forekomme kombinasjoner av inngangsdata og/eller sekvenser av inngangsdata som programvaren så langt aldri har vært utsatt for. Dette betyr at selv om en programvare har fungert i årevis uten at det har oppstått farlige situasjoner, så kan man ikke utelukke at fremtidige farlige situasjoner vil kunne opptre.

Innføring av KI-aktiverte programvarefunksjoner kan forsterke utfordringene beskrevet ovenfor ytterligere, men vil i mange tilfeller ikke endre situasjonen fundamentalt. Det at KI fører til at systemer kan havne i uønskede tilstander vil i mange tilfeller bare være et spesialtilfelle av at andre programvarerelaterte problemer også kan skape uønskede tilstander. Det betyr at de strategiene man anvender for å håndtere denne problematikken også vil være relevant for KI. Dette er diskutert i neste kapittel.

5.2.1 Eksisterende strategier for å håndtere risiko knyttet til programvare

I et overordnet perspektiv kan strategiene for å håndtere programvarerelatert risiko som er relatert til Helse, Miljø og Sikkerhet beskrives som følger:

1. **Bruk av uavhengige kontrollfunksjoner** Det er mange typer operasjoner der det er mulig å stoppe en individuell kontrollfunksjon om nødvendig, men hvor en total nedstengning/stans av prosessen som styres ikke representerer sikker tilstand. Eksempler innenfor olje og gass vil være dynamiske posisjoneringssystemer og ballasteringssystemer brukt på borerigger brukt til havs. For slike typer operasjoner er bruk av uavhengige kontrollfunksjoner vanligvis nødvendig for å kunne opprettholde sikker tilstand i tilfelle kontrollfunksjoner som brukes normalt havner i en uønsket tilstand. Nedenfor diskuteres to forhold som er viktig å vurdere når denne strategien benyttes:

    a. **Manuell aktivering av uavhengige kontrollfunksjoner.** I mange tilfeller finnes ikke uavhengige funksjoner som automatisk er i stand til å detektere alle former for uønskede tilstander uansett årsak. I slike tilfeller vil operatørs evne til å forstå at det har oppstått en uønsket tilstand basert på den totale mengden tilgjengelig informasjon være avgjørende for sikkerheten, siden den uavhengige kontrollfunksjonen i slike tilfeller må aktiveres manuelt. I hvilken grad uavhengige kontrollfunksjoner representerer et realistisk alternativt vil også være avhengig av om operatøren evner å aktivere den tidsnok, og dersom den må opereres manuelt, om dette vil være mulig under alle operasjonelle forhold.

    b. **Uavhengighet mellom rådgivende applikasjoner og kontrollfunksjoner.** I noen operasjoner brukes rådgivende applikasjoner som inngangsdata til operatørers beslutningstaking. KI benyttet i slike rådgivende applikasjoner vil kunne påvirke hvordan operatører bruker kontrollfunksjoner til å kontrollere en prosess, og uønskede resultater fra den rådgivende applikasjon kan derfor føre til at prosessen som blir kontrollert havner i en uønsket tilstand. Deteksjonen av tilstanden vil typisk skje gjennom kontrollfunksjonens sensorer, og kontrollfunksjonen vil også bli forsøkt brukt til å bringe prosessen til en sikker tilstand. I slike tilfeller vil den rådgivende applikasjon og kontrollfunksjonene være uavhengige av hverandre rent teknisk, og kontrollfunksjonen vil kunne være en del av en uavhengig barriere. Men dersom det ikke kommer alarmer fra kontrollfunksjoner, vil deteksjon av at prosessen som blir kontrollert har havnet i en uønsket tilstand være avhengig av at operatør har svært god forståelse av prosessen som blir kontrollert.

2. **Uavhengige sikkerhetsfunksjoner**, som har en høyere autoritet enn kontrollfunksjonene, stenger ned prosessen som blir kontrollert når dette er nødvendig for å opprettholde sikker tilstand. Se også «aktive sikkerhetsfunksjoner» som beskrevet i Havtil's veiledning til § 8 i innretningsforskriften /4/. Innenfor olje og gass blir slike sikkerhetsfunksjoner typisk realisert v.h.a. standardiserte maskinvare og programvarekomponenter sertifisert for bruk i sikkerhetskritiske systemer i henhold til krav i IEC 61508 /56/. Det finnes to varianter av denne strategien, og disse er vesensforskjellige når det kommer til hvor krevende det er å demonstrere at løsningen er sikker.

    a. **Automatisert aktivering av sikkerhetsfunksjoner.** I mange systemer er sikkerhetsfunksjonene fullautomatiserte slik at de utløses basert på avlesninger fra egne sensorer som måler parametere som trykk, temperatur, belastninger etc. Hvis verdien av slike parametere er tilstrekkelig til å beskrive det sikre operasjonsområdet til prosessen som styres, kan det være mulig å argumentere for at KI i kontrollfunksjonen ikke vil påvirke sikkerheten negativt, siden sikkerhetsfunksjonen vil aktiveres dersom KI-algoritmen skulle drive noen av parameterne ut av sikkert område. For å unngå unødige nedstrendinger vil det imidlertid være viktig at KI-algoritmen blir trent til å ikke utfordre grensene som vil trigge sikkerhetsfunksjonene.

    b. **Manuell aktivering av sikkerhetsfunksjoner.** Hvis det er uønskede tilstander i den primære kontrollfunksjonen som ikke kan oppdages av uavhengige sensoravlesninger som beskrevet ovenfor, kan aktivering av sikkerhetsfunksjonen være avhengig av at operatøren oppdager at det er noe galt basert på den totale mengden tilgjengelig informasjon. Hvis det er tilfelle, blir bildet mye mer

komplekst, siden dette reiser en rekke spørsmål knyttet til situasjonsforståelse og menneskelige faktorer. I en slik situasjon vil det typisk være nødvendig med en detaljert risikoanalyse av kontrollfunksjonen for å identifisere relevante uønskde tilstander, og hvordan disse kan observeres for mennesket dersom det ikke kommer alarmer fra den kontrollfunksjonen. Man må også vurdere om det er realistisk at operatøren klarer å aktivere sikkerhetsfunksjonen før det er for sent. I en slik setting kan innføringen av en KI-algoritme i kontrollfunksjonen ytterligere utfordre operatørens situasjonsforståelse og dermed gjøre sikkerhetsdemonstrasjon enda vanskeligere.

3. **Operasjonelle restriksjoner** brukes til å redusere de verste effektene dersom uønskede tilstander i programvaren skulle eskalere til en farlig hendelse. I mange tilfeller er dette den enkleste måten å redusere risiko for skade på mennesker. Et typisk eksempel er operasjoner der mennesker ikke får lov å være i nærheten av farlige sone når operasjonen foregår. Denne måten å håndtere risiko på vil være høyst aktuelt også ved introduksjon av KI.

4. **Kontrollfunksjoner utvikles til et høyt integritetsnivå.** Når denne strategien må brukes kan også bruk av uavhengige kontrollfunksjoner og operasjonelle restriksjoner være relevante strategier, men disse vil ikke kunne redusere risikoen tilstrekkelig dersom verst mulige uønsket tilstand i kontrollfunksjonen skulle opptre. Dermed må kontrollfunksjonen utvikles til et høyt integritetsnivå, og det er leverandørenes sikkerhetsstyringssystem, herunder kvalitetssystemene, som styrer utvikling av system, maskinvare og programvare, som bidrar med den nødvendige risikoreduksjonen. I IEC 61508 omtales slike funksjoner som «kontinuerlige sikkerhetsfunksjoner».

Den fjerde tilnærmingen er typisk for en del systemer innenfor luftfart, bilindustri, jernbane m.fl. Den er svært krevende både kunnskapsmessig og ressursmessig, noe som gjør at leverandører i disse industriene ofte vil være avhengig av å spre kostnadene sine på et større antall like systemer. Relatert til dette er det viktig å være klar over at dersom KI blir brukt til sikkerhetskritiske systemer for eksempel innfor bilbransjen, så vil algoritmen typisk bli implementert eller integrert av organisasjoner som har kompetanse på å utvikle komplekse kontrollfunksjoner til et høyt integritetsnivå. Det er også viktig å være klar over at applikasjoner som inngår i slike funksjoner kjører på sertifisert maskinvare og programvare. Eksempelvis så vil operativsystemer typisk være en variant av et sanntidsoperativsystem sertifisert for bruk i sikkerhetskritiske applikasjoner. VxWorks og QNX er to av mange eksempler på sanntidsoperativsystemer som har slike sertifiserte varianter. Siden svært få norske bedrifter leverer sikkerhetskritiske kontrollsystemer til eksempelvis fly, jernbane eller biler, er det få personer i Norge som har erfaring med denne strategien, som uten sammenligning er den mest krevende av de fire som er diskutert i dette kapittelet.

I kontrast til dette har Norge en stor leverandørindustri innenfor olje og gass og maritim industri, som er industrier som begrenser seg til kombinasjoner avde tre førstnevnte strategiene. Dette betyr at godkjenning er basert på en antagelse om at kontrollfunksjoner på en eller annen måte vil svikte fra tid til annen, og at operasjonelle begrensninger, og/eller uavhengige sikkerhetsfunksjoner og/eller uavhengige kontrollfunksjoner derfor er nødvendig for å håndtere risikoen knyttet til dette. Denne filosofien er også gjenkjennbar fra Havtils barrierenotat /7/.

I mange tilfeller vil også bruk av en eller flere av de tre førstnevnte strategiene være en langt enklere måte å håndtere KI-relatert risiko enn bruk av den siste der man må demonstrere at en KI-løsning er sikker i isolasjon, men i de tilfellene der barrierer må aktiveres manuelt, kan det være mangel på uavhengige deteksjonsmekanismer, se 1a, 1b, og 2b ovenfor. Bruk av KI kan potensielt gjøre det enda vanskelige å få operatør å ta riktig beslutning, og vanskeligere å overta kontrollen av prosessen manuelt i situasjoner hvor dette er aktuelt.

Økt grad av automasjon betyr dermed at kontrollfunksjoner som i dag ikke anses å være sikkerhetskritiske blir mer og mer kritiske, noe som betyr at man beveger seg mot en gråsone der man som i bil, jernbane og flybransjen ser kritiske komplekse funksjoner som kontinuerlig må virke etter intensjon for at sikkerheten skal være ivaretatt. Dette utfordrer regimet for godkjenning av systemer som er kritiske for sikkerheten.

Det er også en utfordring at så lenge en funksjon ikke er definert som en sikkerhetsfunksjon og leverandørene dermed ikke trenger å demonstrere et Safety Integrity Level (SIL) eller lignende, så står leverandørene ganske fritt til å velge maskinvare og annen type hyllevare som f.eks. CPU kort, operativsystemer, kommunikasjonsprotokoller osv. De står også ganske fritt til å velge metodikk for utvikling verifikasjon og validering utover testing.

5.3 Risikoakseptkriterier og hvordan de kan påvirke risikovurderinger knyttet til KI

Nedenfor vises et eksempel på en risikomatrise, som inneholder akseptkriterier knyttet til kombinasjoner av sannsynlighet for farlige hendelser og deres alvorlighetsgrad. Risiko klassifisert som høy vil typisk anses som uakseptabel, mens middels risiko ofte håndteres basert på det såkalte «As Low As Reasonably Practicable» (ALARP)» prinsippet. Eksempelmatrisen er ikke spesielt streng, men det er likevel utfordrende å påvise tilstrekkelig lave sannsynligheter for farlige hendelser som kan få alvorlige eller katastrofale konsekvenser.

Tabell 5-1 Eksempel på risikomatrise med akseptkriterier.

Probablity of dangerous event per year	P >= x/year	None	Negligible	Minor	Significant	Severe	Catastrophic
Frequent	1.E+00	Low	Medium	High	High	High	High
Probable	1.E-01	Low	Low	Medium	High	High	High
Occasional	1.E-02	Low	Low	Medium	Medium	High	High
Remote	1.E-03	Low	Low	Low	Medium	Medium	High
Very Remote	1.E-04	Low	Low	Low	Low	Medium	Medium
Improbable	1.E-05	Low	Low	Low	Low	Low	Medium

I olje- og gassektoren vil det typiske være operatører, som er ansvarlig for å definere denne typen risikoakseptkriterier. Havtils regelverk krever at slike kriterier er på plass og brukes aktivt, men regelverket vil ikke definere grensene for hva som anses som en akseptabel risiko.

Havtils regelverk er i hovedsak funksjonsorientert, noe som betyr at det fokuserer på å oppnå ønsket utfall, uten å foreskrive nøyaktige tekniske løsninger. Det finnes imidlertid et antall tekniske minimumskrav både i regelverket, Havtils veiledning til regelverket og i de standarder og veiledning som forskriften viser til. Dette bety at selv om de ulike eierne kan ha ulike risikoakseptkriterier, medfører ikke dette at det generelle sikkerhetsnivået når det gjelder standardteknologi som brukes innen petroleumsindustrien, vil være tilsvarende forskjellig. Ulike kriterier for risikoaksept blant eiere kan imidlertid være en faktor som må vurderes når ny teknologi som KI innføres, siden det så langt ikke er laget KI-relaterte veiledere som er spesifikke for petroleumsindustrien.

For sikkerhetskritiske funksjoner, der man må demonstrere et «Safety Integrity Level» (SIL) eller lignende, regner man ikke på sannsynlighet for feil i programvare i selve sikkerhetsfunksjonene. Det man gjør er å stille så strenge krav til prosessene rundt risikoanalyser, spesifisering, design, implementering, verifikasjon og validering at man anser sannsynligheten for farlige ikke-detektert feil i programvare som neglisjerbar i forhold til beregnet sannsynligheten for farlige ikke-detektert feil forårsaket av tilfeldige feil i fysiske komponenter. Denne forenklingen kan åpenbart kritiseres, men i mangel av noe bedre, er den typisk for sikkerhetsstandarder på tvers av industrier, se for eksempel IEC 61508 /56/. De strenge kravene gjelder ikke bare den prosjektspesifikke applikasjonen, men all programvare inkludert operativsystemer, kommunikasjonsprotokoller, drivere etc. Kravene blir strengere jo høyere integritet man skal demonstrere, og går langt utover det som er vanlig for programvare i andre sammenhenger. Dette betyr at dersom man ønsker å benytte KI i en sikkerhetsfunksjon så må man kunne argumentere med at kravene som gjelder for programvare i relevante sikkerhetsstandarder er møtt, noe som vil være en svært krevende.

Innenfor petroleumssektoren stilles slike strenge integritetskrav kun til uavhengige barrierer, og ikke til funksjonene som normalt kontrollerer prosessene som barrierene beskytter. Terskelen for å benytte KI i kontrollfunksjoner vil dermed kunne være lavere enn terskelen for å benytte KI i rene sikkerhetsfunksjoner.

Dersom man i forbindelse med risikoanalyse av en kontrollfunksjon, benytter en risikomatrise som vist ovenfor, må man estimere hvor ofte denne funksjonen kan ende opp i en farlig uønsket tilstand. Kombinasjonen av sannsynlighet og verste mulige konsekvens dersom den uønskede tilstanden skulle lede til en hendelse, vil avgjøre om risikoen klassifiseres som Høy, Middels eller Lav.

Det å gjøre en slik vurdering er vanskelig, både fordi det ikke finnes noen omforent metode for å beregne sannsynlighet for uønskede tilstander i programvare, og fordi antakelsen om at svært strenge krav til utviklingsmetodikk gjør at man kan neglisjere denne sannsynligheten kun vil være relevant for funksjoner utviklet iht. til sikkerhetsstandarder. Bruk av KI i kontrollfunksjoner vil typisk kunne gjøre det enda vanskeligere å komme opp med et slikt estimat.

Det er likevel viktig å være klar over at sikkerhetsstandardene som er identifisert i Havtils veiledning til § 5 i styringsforskriften setter noen begrensing knyttet til estimerte sannsynligheten for farlige uønskede tilstander, og at disse vil kunne være relevante også for funksjoner som er basert på KI.

En av begrensningen gjelder funksjoner som kontrollerer en sikkerhetskritisk prosess og som ikke er utviklet til et SIL eller lignende. For slike funksjoner må man anta at en farlig ikke-detektert feil vil skje oftere enn hvert 10 år i operasjon, og det betyr at man kun kan velge kategoriene «Frequent» eller «Probable» i Tabell 5-1 . For å hevde enda lavere sannsynlighet for ikke-detektert farlig feil, må kontrollfunksjonen enten utvikles til et SIL nivå eller lignende slik det gjøres i luftfart, jernbane og bilindustri, eller så må sikkerheten være ivaretatt av uavhengige barrierer, for eksempel en eller flere uavhengige sikkerhetsfunksjoner og/eller andre typer barrierer.

Havtils regelverk tilsier at det er barrierestrategien som er relevant innenfor petroleumsnæringen, og her er det verd å merke seg at for en barriere som krever menneskelig aktiviering anbefaler standardene at risikoreduksjonsfaktoren (RRF) maksimalt skal settes lik 10. Følgelig vil en barriere som har en avhengighet til operatør kun redusere sannsynlighetsestimatet med ett enkelt nivå i risikomatrisen ovenfor, noe som kan gi for lite risikoreduksjon dersom konsekvensen av en hendelse er høy, og man ikke har andre uavhengig barrierer i tillegg.

5.4 Bruk av KI i sikkerhetsfunksjoner

Innføringen av KI i sikkerhetsfunksjoner, der det er nødvendig å demonstrere et Safety Integrity Level (SIL) nivå, eller lignende, medfører en betydelig bevisbyrde. Sikkerhetsfunksjoner i petroleumsindustrien inngår i barrierer og har en relativt enkel logikk. Det betyr at den ekstra kompleksiteten som bruk av KI medfører, ikke lett kan rettferdiggjøres. Av den grunn fokuserer ikke denne rapporten på innføring av KI i slike systemer. Det kan likevel ikke utelukkes at KI-baserte komponenter kan bli innført på lengre sikt. Eksempelvis kan man tenke seg KI-basert aktivering av sikkerhetsfunksjoner kommer som et tillegg, der man i dag kun har menneskelig aktivering.

For dette temaet henvises det til ISO/IEC TR 5469 «Artificial intelligence—Functional safety and AI systems» /52/.

5.5 Cybersikkerhet

KI og cybersikkerhet kan manifestere seg på forskjellige måter. Hackere kan benytte KI for cyberangrep, noe som gjør truslene mer sofistikerte og effektive (se 5.5.1). Samtidig kan KI brukes til å forsvare mot cyberangrep ved å raskt identifisere og reagere på unormale aktiviteter (se 5.5.2). KI kan også være målet for cyberangrep, for eksempel, ved å manipulere treningsdataene for KI-algoritmen eller gjennom dataforgiftning (se 5.5.3). I tillegg kan KI støtte cybersikkerhet i systemutvikling ved å identifisere sårbarheter i koden og forbedre sikkerhetstiltakene (se 5.5.4).

5.5.1 KI som trussel

Hackere kan utnytte KI for å gjennomføre mer sofistikerte og effektive cyberangrep. Dette skaper en rekke nye utfordringer for cybersikkerhet, blant annet:

- KI kan gjøre det mulig for hackere å automatisere angrep og utføre dem i stor skala.
- KI kan brukes til å lage avanserte phishing-angrep som er vanskeligere å oppdage, fordi den kan analysere store mengder data for å lage mer overbevisende svindelmeldinger som er tilpasset individuelle mål.
- Hackere kan benytte KI til å utvikle nye evasjonsteknikker som omgår tradisjonelle sikkerhetssystemer. For eksempel kan KI brukes til å teste forsvarssystemer for å identifisere og utnytte sårbarheter.

En KI-basert tilnærming kan også lære og tilpasse seg på en måte som overgår menneskelige angripere. Dette kan forbedre evnen til å gjennomføre alle disse alternativene.

5.5.2 KI som forsvar

KI kan brukes til å forbedre cybersikkerhetstiltak på flere måter. KIs evne til å behandle store datasett raskt gjør det mulig å identifisere unormale aktiviteter som kan indikere en cybersikkerhetstrussel på et mye tidligere stadium enn tradisjonelle metoder. Gjennom kontinuerlige læringsmekanismer kan KI-løsninger også tilpasse seg nye og utviklende trusler raskere enn statiske, regelbaserte systemer.

Videre kan KI-drevne sikkerhetsprotokoller iverksette nyanserte og dynamiske responser på oppdagede trusler, noe som effektivt kan redusere potensielle konsekvenser. Disse tilpasningsdyktige tiltakene kan inkludere:

- KI-systemer kan automatisk identifisere og klassifisere nye trusler ved å sammenligne med et bredt spekter av kjente mønstre og anomalier.
- Ved å bruke avanserte analyser og prediktiv modellering kan KI forutsi fremtidige angrep basert på historiske data og aktivitetstrender.
- KI kan utvikle automatiserte forsvarstaktikker som tilpasser seg fiendtlige angrep i sanntid, noe som minimerer skadeomfanget og gjenopprettingstiden.

5.5.3 Angrep på KI

KI-systemer kan også åpne for nye angrepsflater. For eksempel kan KI-algoritmer vise seg sårbare for manipulering. Den avhengigheten maskinlæringssystemer har til data, presenterer muligheter for ulike typer datamanipuleringsangrep hvor manipulerte data kan villede KI-systemer.

I tillegg vil KI-systemer være sårbare for de samme typer villede handlinger som annen programvare, inkludert hacking, datavirus, jamming av kommunikasjon, og sabotasje utført av mennesker med fysisk tilgang. Denne generaliserte sårbarheten understreker behovet for omfattende sikkerhetstiltak for å beskytte KI-systemer på alle nivåer.

Eksempler av angrep på KI er:

- Manipulering av treningsdata: Hackere kan manipulere treningsdataene som KI-algoritmer er avhengige av, og dermed påvirke resultatene og oppførselen til KI-systemer. Dette kan gjøre KI-systemene ineffektive eller få dem til å utføre skadelige handlinger.
- Dataforgiftning (data poisioning): Ved å forurense treningsdatasettet kan angripere sikre at KI-modeller lærer feilaktige mønstre, noe som kan føre til feilaktige beslutninger og handlinger.
- Model inversion attacks: Angripere kan rekonstruere sensitive treningsdata ved å analysere modellens respons, noe som kan kompromittere personvernet og sikkerheten til dataene.
- Adversarial attacks: Ved å lage subtile endringer i input-dataene kan angripere mislede KI-modellen til å gjøre feil klassifiseringer eller beslutninger.

5.5.4 KI som støtter cybersikkerhet i systemutvikling

KI kan bistå utviklere og sikkerhetsanalytikere på flere måter. KI-verktøy kan brukes til automatisk analyse av kildekode for å identifisere sårbarheter, inkludert å finne feil som kan utnyttes av hackere eller sårbarheter som ikke var opplagt under manuell koding. KI-systemer kan også overvåke programvare for feil og automatisk logge disse i feil sporings-systemer. Dette forbedrer effektiviteten til utviklerteamene og reduserer tiden det tar å løse problemer.

KI kan brukes til kontinuerlig skanning av systemer og nettverk for eksisterende og nye sårbarheter, og kan foreslå rettetiltak for å forbedre sikkerheten. KI-baserte løsninger kan også benyttes til penetrasjonstesting (pen-testing) for å etterligne cyberangrep og identifisere sikkerhetshull før de blir utnyttet av faktiske angripere.

5.6 Eksempler på risikofaktorer spesifikt knyttet til KI

Dette kapittelet ser på noen risikofaktorer som er spesifikke for KI. Disse er viktige å ta hensyn til i forbindelse med risikoanalyser der man vurder i hvilken grad en operatør eller en uavhengig programvarebasert funksjon vil være i stand til å detektere at resultater fra KI er feil eller ikke er formålstjenlige.

De er også viktige å ta hensyn til når man skal designe KI-løsninger som i seg selv er mest mulig robuste, og de illustrer også at KI baserte systemer typisk vil kreve mer oppfølging i driftsfasen enn systemer som ikke benytter KI.

I tillegg til faktorene som er nevnt i dette kapittelet, er det et antall risikoer som er diskutert i andre deler av rapporten:

- Kapittel 5.2 og 5.3 ser på risikoer knyttet til at implementering av KI som oftest skjer i programvare som ikke blir utviklet til et høyt integritetsnivå, noe som skaper et stort behov for uavhengige deteksjonsmekanismer.
- Kapittel 5.5.3 ser på hvordan KI kan skape nye angreps-flater for cyberangrep.
- Kapittel 5.7 ser på samspillet mellom mennesker og KI, og diskuterer blant annet risikoer knyttet til manglende forklarbarhet, transparens, og operatørens evne til å overvåke det KI-baserte systemet.
- Kapittel 6.2 gir en kort beskrivelse av EUs forordning om KI, og tar opp risikoen knyttet til at et KI-system kan bli benyttet til et annet og mer kritisk formål enn det det opprinnelig var lagd for.

5.6.1 Forringelse av modeller

Modellforringelse oppstår når de forholdene en KI-modell ble trent under, gradvis endrer seg over tid, eller når uventede variabler introduseres i det operasjonelle miljøet. Forringede modeller kan føre til upålitelige prediksjoner eller anbefalinger, som igjen kan ha direkte implikasjoner for operasjonell sikkerhet og effektivitet.

Årsaker til modellforringelse kan være:

- De operasjonelle forholdene kan endre seg betydelig over tid. Dette kan være forårsaket av fysiske endringer i utstyr eller ressurser, endringer i arbeidspersonalets kompetanse eller modifikasjoner i operasjonsprosedyrer.
- KI-modellen kan bli utsatt for data den ikke ble trent på — med ny eller uventet data — noe som kan redusere dens nøyaktighet eller anvendelighet.
- Nye teknologiske fremskritt kan også innføre variabler som ikke ble vurdert under den opprinnelige modelltreningsfasen.

Vi kan forebygge og håndtere modellforringelse på forskjellige måter:

- Periodisk evaluering av modellens ytelse i lys av nye data og endrede forhold, kan gi tidlig identifisering av forringelse. Dette bør gjøres som en del av regelmessig vedlikehold der man om nødvendig, kjører ny opplæring av modellen.
- Implementering av systemer for adaptiv læring der modellen kontinuerlig justeres basert på ny informasjon og feedback kan bidra til å motvirke effekten av modellforringelse.

- Ved å øke forståelsen av hvordan modeller genererer data, kan man lette identifiseringen av når og hvorfor en modell begynner å forringe.
- Automatisert regelbaserte vurderinger av modellens robusthet mot varierte data og forhold kan også spille en rolle i å forutse potensielle svakheter.

Det er viktig at drifts-organisasjonen har kunnskap og verktøy for å kjenne igjen tegn på modellforringelse. Dette inkluderer forståelse for modell-usikkerhet, forutsigbar pålitelighet og evnen til å overstyre KI-baserte beslutninger ved behov.

5.6.2 Skjevhet i data og informasjonsmodeller

Data- og modellskjevhet kan materialisere seg på flere måter, og det kreves systematisk innsats for å identifisere og styre disse skjevhetene effektivt. Dette inkluderer:

- Feedback loop skjevhet: Potensielle positive feedback-sløyfer hvor modellens utdata påvirker etterfølgende datainnganger. Ved å implementere kontroll- og balansetiltak i modellutdata, kan man forebygge snøballeffekten av initielle skjevheter.
- Historisk skjevhet: Det er viktig å anerkjenne og korrigere skjevhet som er innebygd i historiske data for å sikre at KI-modeller ikke viderefører tidligere unøyaktigheter.
- Induktiv skjevhet: Refererer til en samling av antagelser, enten uttrykt direkte eller implisert, som en læringsalgoritme baserer induksjonsprosessen på. Inkorporering av domenekunnskap i KI-modeller hjelper med å motvirke induktiv skjevhet og gir en strukturert måte å introdusere nyttige antagelser på.
- Målingsskjevhet: Verifisering av datainnganger bidrar til å dempe skjevhet som stammer fra feilaktige målinger eller datainnsamlingsmetoder.
- Representasjonsskjevhet: Sikrer et mangfold i datainnsamlingsprosesser for å beskytte mot skjevhet som kunne oppstå fra over- eller underrepresentasjon av visse mønstre eller grupper.

Det å motvirke disse skjevhetene krever en kontinuerlig prosess som må inkorporeres i et rammeverk for styring av data.

5.6.3 Forringelse i datakvalitet

Data danner grunnlaget for bygging og trening av alle KI-modeller. Derfor spiller kvaliteten på data en kritisk rolle for modellens ytelse, pålitelighet og robusthet. Datakvalitet innebærer aspekter som tilstrekkelighet, fullstendighet, nøyaktighet og relevans, i tillegg til fravær av skjevhet.

Utfordringen i mange KI-prosjekter er ikke nødvendigvis mangel på data, men tilgjengeligheten av høykvalitetsdata som er representative for de reelle operasjonelle forholdene. Ofte har man store mengder data for den nominelle operasjonen av et system, men mangler data for ulykkes-tilstander, feil-tilstander, eller anormale drifts-tilstander som man ofte er interessert i å forutsi eller detektere ved hjelp av KI.

Noen eksempler på problemer knyttet til datakvalitet (se ISO 8000) er som følger:

- Mangelfull eller ufullstendig data
- Støy i data
- Feil metadata eller feil data klassifisering
- Inkonsekvent eller upålitelig datainnsamling
- Feilaktig eller irrelevant data
- Data-aldring
- Manglende representativitet
- Skjulte skjevheter i data
- Datafragmentering

Mangel på relevant data fra det virkelige liv, gjør at man i noen tilfeller bruker både virkelig data og syntetiske data produsert ved hjelp av KI for å trene KI. Dersom man gjør dette iterativt, vil mer og mer av treningen i bli basert på syntetiske data og det er en fare for at resultatene konvergerer mer og mer og blir ubrukelige, noe som kalles for modellkollaps. Bruk av syntetiske datasett som er laget for å etterligne statistiske egenskaper som typisk finnes i datasett fra den virkelige verden, kan imidlertid også brukes for å motvirke problemet /71/.

5.6.4 Overtilpassing

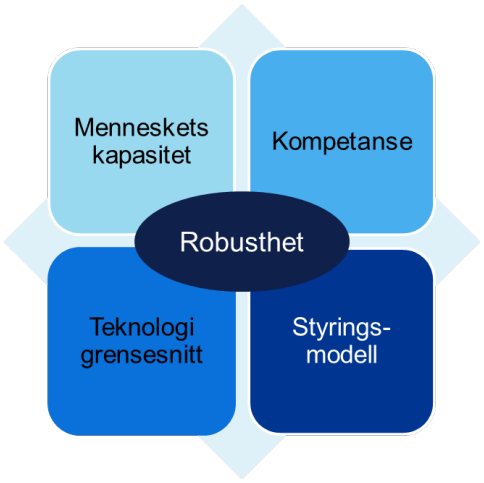
Denne risikoen handler om overtilpassing til treningsdata, hvor en KI-modell tilpasser seg for godt til treningsdata som inkluderer støy og detaljer som ikke er relevante. Dette gjør at modellen presterer godt basert på treningsdata, men presterer dårlig på nye data.

5.7 Samspill mellom mennesker og kunstig intelligens

Moderne KI-aktiverte systemer bruker vanligvis probabilistiske modeller og maskinlæringsalgoritmer som er trent på omfattende datasett for å identifisere mønstre i data. Selv om disse systemene er sterke på mønster-gjenkjenning, er ulempen deres at forholdet mellom input- og outputparametere kanskje verken er helt forståelig eller forutsigbart for brukeren /72/ /73/. Siden KI-baserte systemer «ikke vet hva som er mulig i verden», er det fortsatt en betydelig utfordring å utvikle pålitelige og forutsigbare systemer /73/. Gitt disse begrensningene er slike systemer foreløpig ikke tilstrekkelig egnet til å uavhengig håndtere nye og komplekse situasjoner, noe som krever nøye tilsyn og styring /74/. Dette betyr at menneskelig tilsyn i overskuelig fremtid sannsynligvis vil forbli avgjørende for å overvåke systemytelse, veilede operasjoner og sikre at ønskede resultater blir oppfylt /75/. Men på grunn av deres probabilistiske natur er disse systemene vanskeligere å overvåke og forutsi enn «tradisjonelle» kontrollsystemer /76/. Som et resultat kan operatørene finne det utfordrende å forstå og vurdere atferden deres /75/ /87/. Som et resultat utfordres deres evne til å overvåke.

I denne sammenheng er tillit et relevant psykologisk begrep å vurdere når en designer for samspill mellom mennesker og automasjon /89/. Her defineres tillit som «holdning om at en agent vil bidra til å oppnå et individs mål i en situasjon preget av usikkerhet og sårbarhet» /88/ (s. 51). Tillit er viktig i interaksjon mellom mennesker og KI-systemer ettersom mennesker har en tendens til å bruke systemer de stoler på og avvise systemer de ikke stoler på. Det kan imidlertid være for lite eller for mye tillit til systemet. Mennesker kan bli altfor kritiske til et system og dermed velge å ikke bruke det i det hele tatt («disuse»), eller man kan bli ukritiske og stole for mye på systemet («misuse») /84/. Følgelig, når en designer for tillit i KI-aktiverte systemer, må disse bygges for å være robuste og pålitelige. Videre må disse systemene være i stand til å støtte menneskelig beslutningstaking og tillate intervensjon når det er nødvendig. Derfor må innsats rettes mot å designe systemer for å støtte effektivt samspill mellom mennesker og automasjon. Figur 5-3 illustrerer aspektene som bør vurderes for å designe og operere et avansert og robust system.

- **Menneskets kapasitet** – Vurdere operatørens kognitive belastning, riktig situasjonsforståelse, og evne til å respondere på en uønsket situasjon.
- **Teknologi grensesnitt** – Brukergrensesnittet utformet slik at operatøren får riktig og relevant informasjon over operasjonen og systemets tilstand.
- **Kompetanse** – Kvaliteten på brukeropplæringen med fokus på å kunne bruke systemet til å utføre tiltenkt operasjon, kunne overvåke og forstå systemet, og utføre riktige handlinger ved feiltilstander eller utrygge situasjoner.
- **Styringsmodell** – En moden styringsmodell for utvikling, drift og vedlikehold av systemet, samt oppfølging av brukerne. Dette inkluderer roller og ansvar, godkjente prosedyrer og oppfølging fra ledelsen.

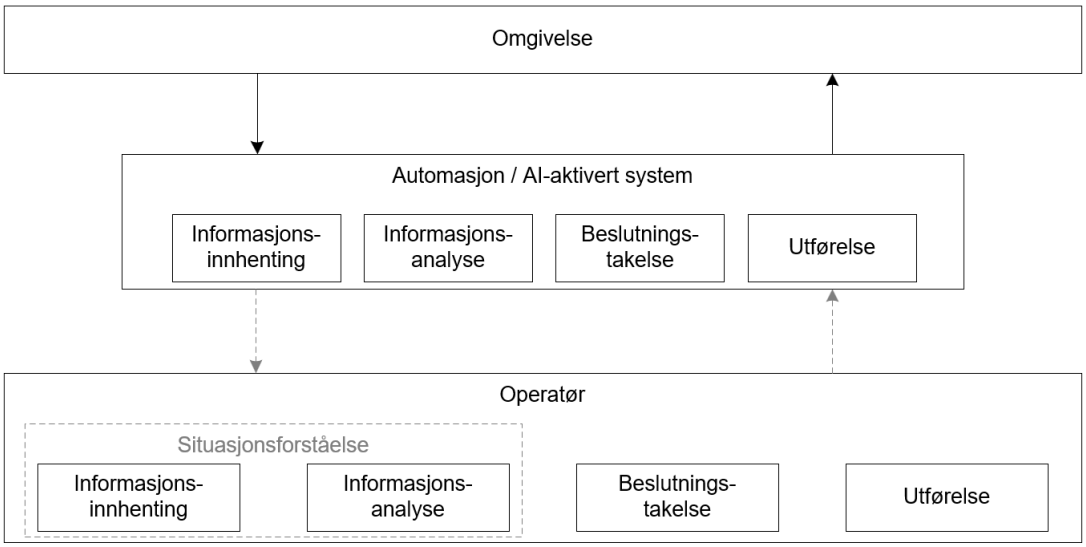


Figur 5-3 Aspekter for et robust menneske-sentrert design.

5.7.1 Samspill mellom menneske og automasjon

Samspillet mellom mennesker og automatisering har lenge vært et forskningstema. Et mål med automasjonssystemer er ofte å erstatte "upålitelig" menneskelig atferd med "pålitelig" automatisering. Dessverre kan det være et avvik mellom systemdesignernes intensjon med systemet og det implementerte systemet, og det kan også hende at det automatiserte systemet ikke lar seg automatisere så mye som ønskelig. For eksempel kan et sikkerhetskritisk system mangle evne til å detektere enkelte uønskede tilstander uavhengig av hva som har forårsaket dem, og dermed vil ikke automatisk overgang til sikker tilstand alltid være mulig. Dette etterlater den menneskelige operatøren, som har i oppgave å overvåke det automatiserte systemet, med (et vilkårlig sett med) gjenværende oppgaver, med lite støtte for å utføre disse /77/. Paradoksalt nok, kan slik (upassende) utforming og bruk av automatisering faktisk forverre, snarere enn å eliminere, menneskets "upålitelighet" ettersom mennesker ikke lenger er en aktiv del av systemets informasjonsbehandling- og beslutningsprosess.

Mennesker som er fjernet fra informasjonsbehandlings- og beslutningssløyfen kan finne det utfordrende å bygge og opprettholde situasjonsforståelse og kan bli "ute av loopen" («out-of-the-loop» (OOTL) se Figur 5-4) /78/. Forskning har vist at dette fører til en rekke utfordringer med menneskelige ytelse og utfordrer deres evne til adekvat å utføre overvåking. Disse inkluderer et ukritisk forhold til systemet (antakelsen om at "alt er bra") /79/ /80/ automatiserings-bias (antakelsen om at systemet sannsynligvis er riktig) /81/, redusert årvåkenhet (på grunn av utarming av mentale ressurser), og misbruk av systemet (på grunn av over- eller underavhengighet av systemet) /84/. Til slutt, når automatisering mislykkes, kan det hende at operatøren ikke er helt oppdatert med den nåværende tilstanden til systemet, noe som resulterer i urimelig høy arbeidsbelastning i et forsøk på å gjenopprette kontroll /85/. Dette er spesielt aktuelt for systemer med høy grad av automasjon or autonomi.



Figur 5-4 Mangelfull innsyn i systemets informasjonsprosessering kan føre til en redusert situasjonsforståelse og at operatøren blir «ute av loopen» /78/.

Ovennevnte automatiseringsrelaterte utfordringer er like relevant for KI-baserte systemer /111/ samtidig som det finnes det nye utfordringer som er spesifikk for KI-baserte systemer /86/. For eksempel, det vil bli utfordrende å forstå hvordan et system fungerer når den selv kan endre seg over tid. Det vil si, systemer som får hyppige oppdateringer, eller har evne til å lære over tid, vil utfordre dette da systemets adferd ikke vil være like forutsigbare som tradisjonelle systemer. Videre kan beslutningsstøttesystemer bruke store språkmodeller for å utvikle og kommunisere planer. Med tanke på realismen i språkbruk med slike systemer, kan formuleringen av disse planene virke overbevisende for mennesker. Dette kan øke deres tilbøyelighet for å være enig i systemets forslag i stedet for å vurdere det grundig. Dette er spesielt relevant dersom systemer oppfattes som pålitelige, noe som igjen fører til at brukerne tror disse systemene er mer kapable enn de egentlig er /86/. Dessuten kan de miste fokus og engasjere seg i andre oppgaver /92/ /93/. Med tanke på den utbredte bruken av KI-baserte systemer i hele samfunnet for lavrisikoapplikasjoner, for eksempel store språkmodeller tilgjengelig i standard kontorapplikasjoner, kan dette gjøre brukeren ufølsomme for dets mangler. Følgelig kan risikoen forbundet med disse systemene utilsiktet migrere fra lavrisiko- til høyrisiko-applikasjoner.

Disse eksempler illustrerer at personer som har i oppgave å overvåke avanserte og kapable KI-baserte systemer, forventes – men muligens ikke er i stand til – å gripe inn ved behov. Følgelig ender ofte operatørene med å akseptere systemets forslag, det vil si at operatøren blir en trent "knappe trykker" som mister ferdighetene sine for å utføre sine overvåkingsoppgaver tilstrekkelig /94/. Dette betyr at operatøren kanskje ikke kan utøve sin rolle som en uavhengig sikkerhetsbarriere på en god nok måte. I sum er samspillet mellom mennesker og automatisering et intrikat problem – jo mer kapabel og pålitelig systemet blir, jo mindre sannsynlig er det at operatøren tar manuell kontroll over systemet i en kritisk situasjon, og jo større kan konsekvensene av å ikke gripe inn bli /75/.

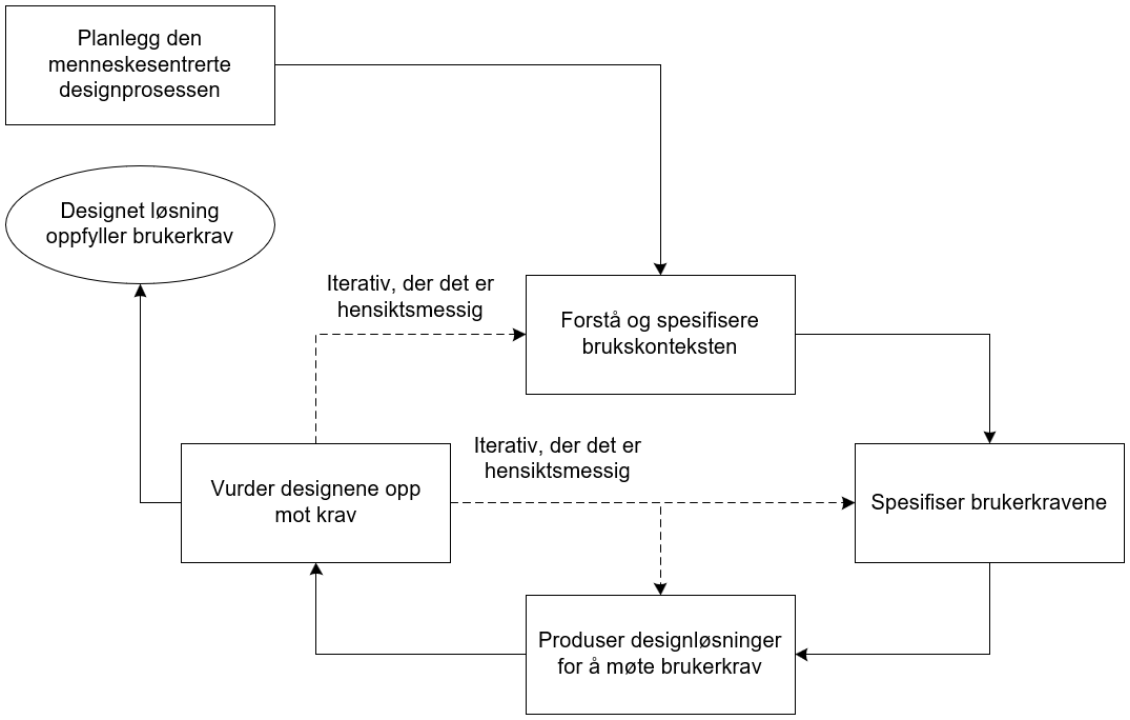
5.7.2 Strategier for å støtte operatørytelse

Det er utviklet en rekke strategier for hvordan man kan støtte samspill mellom mennesker og automasjon /75/ /95/ /96/ /97/. Med tanke på at OOTL problemet er preget av redusert evne til å oppdage systemfeil og utføre oppgaver manuelt når automatiseringen svikter, er en viktig avbøtende strategi å involvere dem på nytt i systemets beslutningsprosess /78/. Dette betyr at når man designer for menneskelig overvåkingsytelse av automatiserte systemer (inkludert KI-baserte), bør det tas hensyn til hva operatøren forventes å gjøre, hvordan de forventes å gjøre det, og hvilken informasjon de trenger for dette. På denne måten kan involvering i oppgaven og den menneskelig beslutningsprosessen utformes av menneske-automatiserings teamet.

Menneskesentrert design («Human Centered Design» - HCD) er en tilnærming der funksjoner, oppgaver og menneskelige behov tas i betraktning i systemdesign /95/ /98/. I motsetning til teknologi-sentrert design, vurderer HCD brukerens oppgaver, kontekst og evner for å informere systemdesign, for eksempel for utforming av menneskelige maskingrensesnitt. For å oppnå en HCD-prosess bør følgende hovedaktiviteter utføres (se Figur 5-5) /98/:

- Systemets brukskontekst spesifiseres og forstås,
- Brukerkrav spesifiseres,
- Design-løsninger produseres,
- Løsningen evalueres og testes.

Gjennom en serie iterasjoner, hvor brukerkrav og designløsninger oppdateres og evalueres, produseres en designløsning som skal møte brukerkravene. Tidligere forskning og erfaring fra bruk av denne prosessen for nybygg og modifikasjoner på norsk kontinentalsokkel, har vist at HCD-prosessen er en effektiv metode for å utvikle menneskesentrerte automatiserte systemer /99/ /100/ /101/ /102/. Dette betyr at systemdesignere, som har i oppgave å utvikle menneskesentrerte KI-baserte systemer, har en etablert prosess til rådighet for å utvikle systemer som støtter menneskelig overvåking.



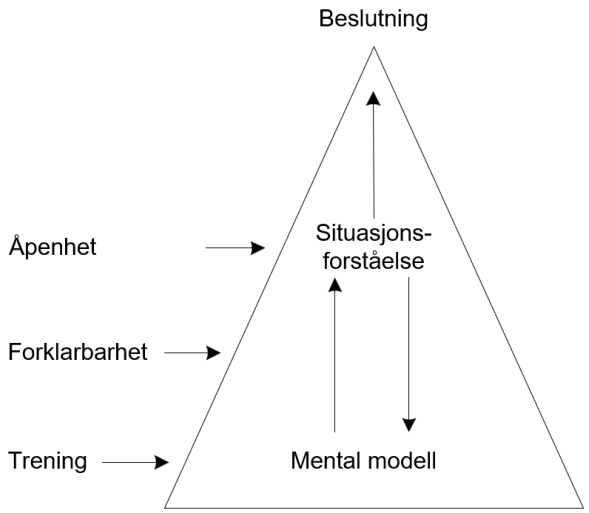
Figur 5-5 Menneskesentrert designprosess /98/.

Ved utforming av KI-basert system bør det legges vekt på menneskets behov for å forstå disse, minimere kompleksiteten deres og gi støtte til situasjonsforståelse. Her er tilstrekkelig og hensiktsmessig tilbakemelding fra systemet et av de sentrale elementene i å støtte mennesker i å skape adekvate mentale modeller av systemet /103/. Hva som teller som "tilstrekkelig og hensiktsmessig tilbakemelding" varierer imidlertid med oppgaven og operasjonskonteksten. For eksempel stiller systemer for tidskritiske operasjoner andre krav sammenlignet med de systemene hvor det er mere tid til å vurdere tilgjengelig informasjon og ta bedre funderte beslutninger. Det vil si at i tidskritiske situasjoner hvor operatører har behov for å ta raske beslutninger, bør informasjonsgrunnlaget og

samhandlings-mulighetene fra systemet være tilpasset operatørens behov. Dette kan innebære å presentere få men likevel essensielle parametere, bruke visuelle signaler eller lydsignaler, og konsise forslag til aksjon.

Ikke-tidskritiske applikasjoner trenger ikke å begrense informasjonsmengden i samme grad som tidskritiske applikasjoner, da operatøren har mere tid for å ta til seg informasjon og vurdere mulige tiltak. Hva som er tidskritisk er svært avhengig av systemets kontekst og valg angående beslutningsmyndighet, funksjonsfordeling, oppgaver og ytelseskrav fordelt mellom systemet og operatøren. Uansett bør systeminformasjon som et minimum samsvare med operatørenes beslutningsstrategier, ved å presentere relevant informasjon på et brukergrensesnitt som sørger for at operatøren har riktig situasjonsforståelse /104/. I tillegg, bør det settes krav til systemets egen evne til å detektere feiltilstander eller økt usikkerhet. Det vil si, hvis det digitale systemet er degradert på en eller annen måte slik at den presenterte informasjonen kan være feil, så må operatøren bli advart slik at de kan ta hensyn til det og få systemet i en sikker tilstand.

Når en operatør har som oppgave å overvåke og samhandle med KI-aktivertbaserte systemer, blir systemets forståelighet og forutsigbarhet viktige. Fordi disse systemene kan endre seg over tid, blir det stadig vanskeligere å trene operatører til å opprettholde nøyaktige mentale modeller. Mentale modeller er viktige mekanismer som mennesker bruker for å forstå hvordan systemer fungerer, hvilken informasjon som anses nødvendig for å tolke systematferden, og opprettholde riktig situasjonsforståelse /91/. Derfor spiller forklarbarhet («explainability») en viktig rolle i å avsløre hvordan systemets algoritmer fungerer, dets pålitelighet, modellytelse, kontekstuell informasjon, risikoforklaringer, nøyaktighet til prediksjoner, og systemintegritet /89/ /105/. Videre spiller åpenhet («transparancy») en rolle i å bygge brukerens tillit til systemet slik at en formålstjenlig avhengighet av systemet fremmes og passende beslutninger om automatiseringsbruk kan tas (se Figur 5-6) /106/ /107/ /108/ /109/.



Figur 5-6 Forklarbarhet og åpenhet i en beslutningssammenheng /74/ /87/.

Når operatører samhandler med systemer som gir anbefalinger eller utfører handlinger som har sikkerhetskritiske konsekvenser, er innsikt i systemets resonnement viktig for effektiv overvåking /109/. Derfor bør slike systemer kunne gi «forklaringer» og være «åpen» om deres beslutninger og handlinger. I denne sammenheng gir forklarbarhet «informasjon om måten logikken, prosessen, faktorene eller resonnementene som systemets handlinger eller anbefalinger er basert på» /74/, s.31). Forklarbarhet bidrar med andre ord til å forstå logikken som brukes av systemet, dets muligheter og begrensninger ved å gi retrospektiv informasjon bak beslutningene. På denne måten støtter forklaringsevnen operatørens forståelse av hvordan systemet fungerer, når det vil fungere, og når det ikke vil fungere /87/.

Åpenhet «gir en forståelse av handlingene til KI-systemet som en del av situasjonsforståelse» /74/, s.31. Åpenhet støtter operatøren ved å gi mulig informasjon om hvordan systemet fungerer i sanntid og informerer om planlagte handlinger i nær fremtid. Dette betyr at åpenhet bidrar til å informere operatøren om systemets situasjonsforståelse, og til å forstå dets beslutninger og handlinger. I sum støtter og opprettholder forklarbarhet de mentale modellene som ligger til grunn for operatørens forståelse av hvordan systemet fungerer, mens åpenhet direkte støtter operatørens situasjonsforståelse av systemet i dets oppgavemiljø ved å gi aktuell og relevant informasjon i sanntid /87/ /102/ /110/.

Merk at det som er beskrevet så langt i dette kapittelet kun utgjør førstelinjeforsvaret dersom en operasjon skulle ende opp i en utrygg tilstand forårsaket av KI. Sikkerhetsfilosofien innenfor petroleumsindustrien er basert på bruk av uavhengige barrierer, og ikke på at enkeltfunksjoner som inngår i kontrollsystemer eller rådgivende applikasjoner, skal utvikles slik at de har en høy integritet. Dette betyr at man ikke kan basere sikkerheten på en antakelse om at det alltid foreligge en alarm dersom en utrygg tilstand oppstår. Man har derfor et behov for å detektere den utrygge tilstanden ved hjelp av informasjon som er tatt frem uavhengig av det KI-baserte systemet.

I noen tilfeller kan slik uavhengig deteksjon skje før informasjon fra KI-systemet blir tatt i bruk, eksempelvis kan operatørene i noen tilfeller sammenligne resultater fra et rådgivende systemer som benytter KI, med resultater fra andre rådgivende systemer som ikke benytter KI. I andre tilfeller vil uavhengig deteksjon kun være mulig basert på observasjon av data fra prosessen som blir kontrollert.

I noen tilfeller vil uavhengig barrierefunksjoner kunne aktiveres automatisk basert på måling av helt spesifikke prosessparametere, mens i andre tilfeller vil et menneske måtte ta beslutningen om aktivering basert på den totale mengden tilgjengelig informasjon. I slike tilfeller vil evnen til å forstå at prosessen som blir kontrollert har havnet i en utrygg tilstand kunne være viktig. Slik forståelse fordrer enten at systemet som kontrollerer prosessen er uavhengig av det KI-baserte systemet, og dermed kan gi en alarm basert på prosesstilstand, eller at operatøren evner å dra konklusjoner basert på forskjellige målinger av prosessen som blir kontrollert. Dersom operatøren selv skal kunne dra konklusjoner basert på rådata fra målinger, krever dette en dyp innsikt i hvordan prosessen som blir kontrollert fungerer. I de fleste tilfeller vil en slik innsikt være langt mer verdifull enn innsikt i hvordan KI-algoritmen fungerer.

Uansett kunnskapsnivå vil det kunne være vanskelig å identifisere avvikssituasjoner manuelt, dersom avvikene fra normal prosess er relativt små, noe som vil kunne være et problem dersom man også operer med små marginer mot utrygge prosesstilstander. I tillegg så vil ikke måledata fra prosessen som blir kontrollert alltid representere uavhengig informasjon, siden det kan være muligheter for felles-feil i systemet som kontrollerer prosessen, f.eks. i programvare. Se også kapittel 4.3, 5.2 og 5.3 når det gjelder bruk av barrierer og problemstillinger knyttet til uavhengig deteksjon.

Industrien bør derfor utforske mulighetene for ytterligere automatisering av uavhengige barrierefunksjoner.

5.7.3 Mot menneske og KI samarbeid

Det er potensielt store fordeler å oppnå ved å kombinere mennesker og KI-baserte systemer, men implementeringen av slike løsninger bør styres nøye. Utviklingen innen KI er også rask og kunnskapen er i stadig utvikling. På samme måte innebærer dette at forskningsfeltet på samspill mellom mennesker og KI-systemer er under konstant utvikling. Til tross for innsatsen for å lage en oversikt over dagens kunnskapsstatus innenfor feltet, er det mange usikkerhetsmomenter og uløste problemstillinger /74/. Eksisterende kunnskap om Human Factors-metoder er anvendelige og nyttige for KI-baserte systemer, men det er et presserende behov for å videreutvikle kunnskap på tvers av en rekke dimensjoner om hvordan mennesker og KI-baserte systemer kan effektivt samarbeide. Derfor bør innsatsen rettes mot å sikre at mennesker er i stand til å forstå og forutsi atferden til systemet, metoder for å forstå når man kan stole på systemet og når man ikke kan stole på systemet, for på den måten kunne ta gode beslutninger basert på tilgjengelig informasjon, og samtidig ha evnen til å utøve kontroll over systemet på en meningsfull og tidsriktig måte /74/.

5.8 KI-generert kode

KI-generert kode omfatter automatisert skriving av programvarekode ved hjelp av maskinlæringsmodeller og algoritmer. En måte å arbeide på vil være at en utvikler lar KI lage et utkast som vedkommende så forbedrer gjennom manuell koding, men det vil også kunne være aktuelt å bruke KI-generert kode uten å modifisere den på noen måte.

Når det gjelder krav til verifikasjon og validering av slik kode er det naturlig å sammenligne med hvordan kode som er automatisk generert fra modeller blir håndtert. Slik generering har vært vanlig praksis, også innenfor utvikling av sikkerhetskritiske systemer, i forskjellige industrier i flere tiår, eksempelvis i fly og bilindustrien. For sikkerhetskritisk kode finnes det to modeller man kan følge:

1. Koden blir verifisert og validert slik man ville ha gjort det om den var skrevet av et menneske. Dersom dette er tilfelle, blir ikke KI-generert kode håndtert annerledes enn kode skrevet av et menneske.
2. Man velger å stole på generatoren, og lar være å gjennomføre alle V&V-aktiviteter som ville blitt utført for kode laget av mennesker. Eksempelvis vil manuelle inspeksjoner kunne fokusere på modellen som man genererer koden fra, mens man lar være å gjennomføre manuell kodeinspeksjon og andre implementasjonsorienterte V&V aktiviteter.

Dersom man følger modell nummer to ovenfor, vil dette for sikkerhetskritiske systemer utløse krav om kvalifisering av verktøyet som genererer koden. For høykritiske kode vil det antakelig være umulig å få et slikt verktøy kvalifisert i dag.

På samme måte vil verktøy som automatiserer deler av V&V prosessen, eksempelvis gjennom å generere kjørbare sett med tester, måtte gå gjennom en kvalifisering, dersom det som blir produsert ikke blir verifisert og validert av et menneske.

Innenfor luftfarten finnes det en egen veiledning for modellbasert utvikling og verifikasjon av programvare. Denne beskriver hvilke deler av standard veiledningen som bortfaller og hva som kommer i tillegg ved modellbasert utvikling og/eller modellbasert verifikasjon.

Når det gjelder KI utviklet for bruk i sikkerhetsrelaterte systemer innenfor olje og gass vil det være naturlig å følge en tilsvarende modell. Siden det i dag ikke finnes noe regime for å kvalifisere verktøy for programvare som ikke har SIL krav på seg, vil det enkleste være å gjennomføre de samme V&V-aktivitetene som blir gjennomført for manuelt genert kode. Det vil imidlertid kunne være en utfordring at kravene til V&V ikke er spesielt strenge. Dette gjelder spesielt for rådgivende systemer.

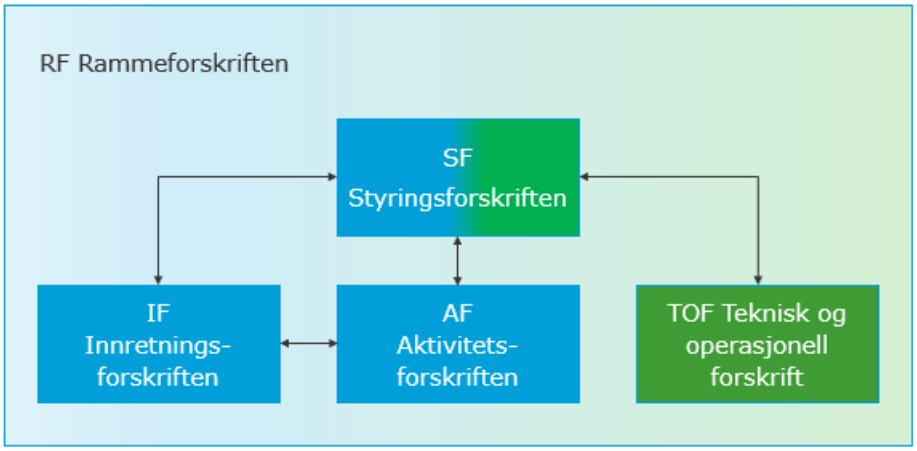
5.9 KI-genererte dokumenter

Det må forventes at KI vil bli brukt i forbindelse med produksjon av en lang rekke forskjellige typer dokumenter. Dette gjelder alt fra arbeidsprosedyrer til tekniske spesifikasjoner, og test prosedyrer. Som i tilfellet kildekode ovenfor vil det være viktig at personell med relevant ekspertise gjennomfører manuell verifikasjon og validering av slike dokumenter.

6 REGULATORISKE OG STANDARDISERINGSKRAV

6.1 Havtils regelverk

Havtils regelverk for olje- og gassvirksomhet inneholder fem forskrifter: *Rammeforskriften* /1/ og *Styringsforskriften* /2/ inneholder overordnede prinsipper for risikoreduksjon og risikokontroll. *Aktivitetsforskriften* /3/, *Innretningsforskriften* /4/, samt *Teknisk og Operasjonell Forskrift* /5/ beskriver drifts- og designprinsipper. Som illustrert i Figur 6-1 nedenfor er Rammeforskriften og Styringsforskriften overordnet de tre andre. Denne forskriften gjelder helse, arbeidsmiljø og sikkerhet ved landanlegg.



Figur 6-1 Regelverkets struktur, fra /6/.

Egne veiledninger til forskriftene viser hvordan bestemmelser i en forskrift kan oppfylles. Forskriftene og veiledningene må sees i sammenheng for å få best mulig forståelse av hvordan forskriftskravet skal innfris. Veiledningene viser på enkelte områder til industristandarder, som en anbefalt måte å oppfylle forskriftens krav på. Veiledningene til forskriftene er ikke rettslig bindende, og aktørene kan derfor velge andre løsninger.

Dersom den ansvarlige aktøren velger å benytte den anbefalte løsningen, kan det normalt legges til grunn at forskriftenes krav er oppfylt. Hvis aktøren velger andre løsninger, som for eksempel andre standarder eller selskapsspesifikke prosedyrer, må de kunne dokumentere at den valgte løsningen er minst like god som, eller bedre enn, den anbefalte.

Det er aktørenes ansvar å følge regelverket som i stor grad er formulert som *funksjonskrav*, med fokus på behov og ønskede resultater, fremfor å foreskrive konkrete løsninger. Dette gir stor grad av frihet til aktørene, men krever også en aktiv tilnærming der aktørene må definere egne risikoakseptkriterier og resultatkrav.

Et grunnleggende designprinsipp er at svikt i en komponent, et system eller en enkelt feil ikke skal føre til uakseptable konsekvenser (innretningsforskriften § 5). Hvordan dette skal oppnås er nærmere beskrevet i andre deler av regelverket, for eksempel at et system går inn i eller opprettholder en sikker tilstand dersom et delsystem svikter (innretningsforskriften § 33 og § 34), bruk av flere barrierer/redundans (innretningsforskriften § 33 og § 34), barrierenes uavhengighet (styringsforskriften § 5, innretningsforskriften § 33 og § 34), og evne til å oppdage problemer og kjenne barrierestatus (aktivitetsforskriften § 26, innretningsforskriften § 34a).

I tillegg til veiledningene som er direkte knyttet til regelverket, har Havtil utgitt et notat knyttet til barrierestyring /7/ og et notat knyttet til risikostyring /8/ som gir ytterligere veiledning relatert til paragrafene nevnt ovenfor.

> ***«Barrierer skal komme i tillegg til en sikker og robust løsning»***
>
> *Uansett hvor godt en forsøker å få til en sikker og robust løsning, vil feil, fare- og ulykkessituasjoner inntreffe. Da skal barrierer tre i kraft og bidra til å håndtere slike situasjoner.*
>
> **Fra Barrierestyringsnotatet /7/**

Sett i kontekst av KI brukt i applikasjoner som er sikkerhetsrelatert, så betyr teksten fra barrierenotatet ovenfor at operatører skal gjøre alt man kan for å gjøre en slik KI-løsning sikker og robust, men at man i tillegg må ha effektive uavhengige barrierer.

Notat knyttet til risikostyring legger særlig vekt på at beslutninger må ta hensyn til usikkerhet, føre var prinsippet, og læring. Notat bruker begrepet risikoinformert virksomhetsstyring for å tydeliggjøre at beslutninger ikke bør tas basert på risikoanalyser alene.

> ***«Å ta hensyn til usikkerhet betyr å systematisk lete etter potensielle overraskelser»***
>
> *Ofte konkluderes det med at man kan se bort fra en hendelse på grunn av lav vurdert sannsynlighet. Slike sannsynlighetsvurderinger kan bygge på uriktige eller svake forutsetninger. Å 'ta hensyn til usikkerhet' betyr å systematisk fokusere på dette problemet og lete etter potensielle overraskelser. I dette arbeidet er det spesielt viktig å være oppmerksom på det som er kjent i organisasjonen eller industrien utenfor, men som er ukjent for de som gjør vurderingene (såkalte «ukjente kjente»).*
>
> **Fra «Integrert og helhetlig risikostyring i petroleumsindustrien» /8/**

Som beskrevet i kapittel 5.1, har regelverket også krav om kvalifisering av ny teknologi.

6.2 EU forordningen om kunstig intelligens

Formålet med EUs forordning om KI /50/, er gjennom harmonisert lovgivning for håndtering av KI, å bidra til at EUs indre marked fungerer best mulig, samt å bidra til et høyt nivå av beskyttelse mot negative effekter fra innføring av denne type teknologi.

EU forordningen stiller strenge krav til systemer som klassifiseres som høyrisiko KI-systemer, og reglene som avgjør om et system havner i denne klassen finnes i kapittel III, seksjon I, Artikkel 6 i reguleringen. Se utdrag nedenfor.

> 1. *Irrespective of whether an AI system is placed on the market or put into service independently of the products referred to in points (a) and (b), that AI system shall be considered to be high-risk where both of the following conditions are fulfilled:*
>
>   (a) *the AI system is intended to be used as a safety component of a product, or the AI system is itself a product, covered by the Union harmonisation legislation listed in Annex I;*
>
>   (b) *the product whose safety component pursuant to point (a) is the AI system, or the AI system itself as a product, is required to undergo a third-party conformity assessment, with a view to the placing on the market or the putting into service of that product pursuant to the Union harmonisation legislation listed in Annex I.*
>
> 2. *In addition to the high-risk AI systems referred to in paragraph 1, AI systems referred to in Annex III shall be considered to be high-risk.*
>
> **Utdrag fra kapittel III, Seksjon I Artikkel 6 i EUs forordring om KI /50/**

Punkt 1 i artikkel 6 referer Anneks I som igjen referer til EU lovgivning relevant for typegodkjenning av spesifikke typer av produkter og systemer. Flere av disse systemtypene brukes innenfor petroleumsnæringen, eksempelvis systemer

som må være i henhold til maskindirektivet /69/ og systemer som må være i henhold til direktivet for utstyr brukt i potensielt eksplosiv atmosfære /70/.

Artikkel 6 punkt 2 referer til Anneks III som identifiserer systemer brukt innenfor forskjellige områder i samfunnet, og som vil havne i høyrisiko kategorien dersom de benytter KI. Flere av disse områdene vil kunne være relevante for petroleumsnæringen, f.eks. vil systemer som personalavdelinger bruker, kunne havne i høyrisikokategorien dersom KI benyttes.

Det området som anses mest relevant for denne rapporten og som er nevnt i Anneks III er KI systemer brukt som sikkerhetskomponenter innenfor kritisk infrastruktur, siden systemer som brukes for gassleveranser faller innenfor denne kategorien, som vist nedenfor.

> *Critical infrastructure: AI systems intended to be used as safety components in the management and operation of critical digital infrastructure, road traffic, or in the supply of water, gas, heating or electricity.*
>
> **Fra Anneks III i EUs forordring om KI /50/**

Fokuset på akkurat disse typene av kritisk infrastruktur skyldes at bortfall av leveranser kan utgjøre en fare for befolkningen. Dette er gjenspeilet gjennom at definisjonen av sikkerhetskomponent nedenfor omfatter mer enn det man i petroleumsnæringen vanligvis definerer som sikkerhetssystem.

> *'safety component' means a component of a product or of an AI system which fulfils a safety function for that product or AI system, or the failure or malfunctioning of which endangers the health and safety of persons or property;*
>
> **Fra Artikkel 3 «Definisjoner» I EUs forordning om KI /50/**

DNVs forståelse av de to definisjonen overfor er at både tradisjonelle sikkerhetssystemer, og systemer som kan forårsake bortfall av gassleveranser, vil bli å anse som et høyrisiko KI-system i henhold til KI forordningen dersom de benytter KI.

Artikkel 6 inneholder også noen unntak som vil være relevante, når man gjennomfører en klassifisering av systemer. Dette handler om systemer som faller inn under områdene beskrevet i Anneks III, men som anses å ha lavere risiko fordi systemene skal: utføre enkle prosedyrer, forbedre resultater fra menneskelige aktiviteter, kommer i tillegg til menneskelige beslutningsprosesser, eller kun brukes til forberedende arbeide. Se detaljert beskrivelse nedenfor.

Disse unntakene illustrere at systemer som brukes til rådgivning, planlegging og tilstandsovervåking som beskrevet i kapittel 4.1 vil kunne havne i en gråsone når det gjelder klassifisering i henhold til KI forordningen.

> *By derogation from paragraph 2, an AI system referred to in Annex III shall not be considered to be high-risk where it does not pose a significant risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making.*
>
> *The first subparagraph shall apply where any of the following conditions is fulfilled:*
>
> *(a) the AI system is intended to perform a narrow procedural task;*
>
> *(b) the AI system is intended to improve the result of a previously completed human activity;*
>
> *(c) the AI system is intended to detect decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment, without proper human review; or*
>
> *(d) the AI system is intended to perform a preparatory task to an assessment relevant for the purposes of the use cases listed in Annex III.*

Notwithstanding the first subparagraph, an AI system referred to in Annex III shall always be considered to be high-risk where the AI system performs profiling of natural persons.

Utdrag fra kapittel III, Seksjon I Artikkel 6 i EUs forordring om KI /50/

KI-algoritmer vil typisk være generiske algoritmer som kan brukes til mange forskjellige formål. Reguleringen legger imidlertid stor vekt på at KI-systemer blir laget for et spesifikt formål (intended purpose), og at alt fra risikoanalyser til trening av algoritmer må være tilpasset til dette formålet.

Det er en mulighet for at KI-systemer som ble laget for et spesifikt formål etter hvert kan bli brukt til et annet formål. Kapittel III, seksjon 3 Artikkel 25.1.c. påpeker at dersom en type KI basert system som allerede er i bruk, får et nytt bruksområde, og dette nye formålet nå gjør at det blir klassifiser som et høyrisikosystem, så må alle krav som reguleringen stiller til høyrisikosystemer oppfylles.

Aktører som utvikler og opererer høyrisikosystemer må ha styringssystemer for risikohåndtering og kvalitetssikring på plass, og må bl.a. oppfylle regulerings krav til testing, ytelse, dataforvaltning, dokumentasjon, før systemene kan tas i bruk. Kravene er omfattende og innebærer også at KI-systemene skal være transparente, sporbare og forklarbare for å sikre at beslutninger som tas av KI kan forstås og etterprøves av de som operer systemene.

Foreløpig må den enkelte aktør som ønsker å ta i bruk KI selv konkretisere og svare ut kravene i forordningen i egne styringssystemer. Slik operasjonalisering av høynivå krav er vanligvis arbeidskrevende, men arbeidsbyrden på hver enkelt organisasjon bør kunne reduseres dersom industrien går sammen om å ta frem felles veiledninger.

6.3 Relevante internasjonale standarder

ISO/IEC 5338: Information technology – Artificial intelligence – AI system life cycle processes /51/ gir veiledning om livssyklusstyring av KI-systemer, med vekt på den systematiske og strukturerte tilnærmingen som kreves fra start til dekommisjonering. Den krever strenge kvalitetssikrings- og prosjektstyringsprosesser, for å sikre at KI-systemer er pålitelige, sikre og effektive.

ISO/IEC 42001: Information technology – Artificial intelligence – Management system /55/ legger grunnlaget for å etablere et KI-styringssystem, som skisserer krav til ledelse, planlegging, støtte, drift, evaluering og forbedring. ISO 42001 er bygget over samme mal som ISO 9000, ISO 31000, ISO 27001 m.fl. og man kan sertifiseres i henhold til den.

ISO/IEC TR 5469: Artificial intelligence – Functional safety and AI systems /52/ Dette er en teknisk rapport som har til hensikt å gjøre det mulig for utviklere av sikkerhetsrelaterte systemer å ta i bruk KI-teknologier i sikkerhetsfunksjoner, gjennom å gi en bedre forståelse for: egenskapene som forskjellige KI teknologier har, hva som er risikofaktorene, hvilke metoder innenfor fagområdet funksjonell sikkerhet som vil være relevante, samt mulige begrensninger. Rapporten har mange koblinger til IEC 61508.

ISO/IEC 22989: Informasjonsteknologi – Kunstig intelligens – Begreper og terminologi for kunstig intelligens /53/ gir en oversikt over grunnleggende konsepter og terminologier innen KI. ISO/IEC 22989 dekker aspekter som forklarbarhet, robusthet, risikostyring og livssyklus av KI-systemer.

ISO/IEC 23894: Information technology – Artificial intelligence – Guidance on risk management /54/ fokuserer direkte på risikostyring relatert til anvendelsen av KI-systemer. Denne standarden gir anvisninger for identifisering, vurdering, og håndtering av risikoer knyttet til innsats og bruk av kunstig intelligens i operasjonelle sammenhenger. Det understreker viktigheten av en kontinuerlig prosess for risikovurdering for å oppdage og redusere potensielle trusler mot sikkerheten og påliteligheten til KI-integrerte systemer.

IEC 61508: Functional safety of electrical/electronic/programmable electronic safety-related systems /56/ angir kravene til funksjonell sikkerhet for elektriske, elektroniske og programmerbare elektroniske systemer som brukes til sikkerhetsrelaterte formål. Standarden dekker hele livssyklusen av et system, fra det tidlige konseptstadiet til

dekommisjonering. Det er en tverrindustriell standard som er anvendelig i mange sektorer hvor sikkerhetskritiske systemer er i bruk. Havtils regelverk referer til denne standarden i paragrafer som omhandler sikkerhetsfunksjoner, og det er vanlig at komponenter som brukes i sikkerhetsfunksjoner innenfor petroleumsindustrien er sertifisert iht. til denne standarden.

IEC 61511: Funksjonell sikkerhet – Sikkerhetsinstrumenterte systemer for prosessindustrien /57/ er avledet av IEC 61508 og er spesifikt rettet mot sikkerhetsinstrumenterte systemer (SIS) brukt i prosessindustrier, som kjemiske, petrokjemiske og olje- og gassindustrien. Standarden har en forventning om at alle software og hardware komponenter som benyttes i et SIS har blitt utviklet i henhold til relevante krav i IEC 61508, noe som gjør at IEC 61511 kan ha et systemfokus. Havtils regelverk referer denne standarden i paragrafer som omhandler sikkerhetsfunksjoner

CEN-CLC/JTC 21: CEN og CENELEC har mottatt en standardiseringsforespørsel knyttet til høy-risiko KI systemer fra Europakommisjonen. I denne forbindelse utvikler CEN-CLC/JTC 21 nå europeiske standarder. De europeiske standardiseringsorganisasjonene har opprinnelig frist til slutten av april 2025 for å ferdigstille og publisere disse standardene. Kommisjonen vil deretter vurdere og eventuelt godkjenne standardene, som vil bli publisert i Den Europeiske Unions offisielle journal. Når de er publisert, vil standardene gi en "formodning om samsvar" med EUs AI Act for KI-systemer Dvs. oppfyller man standardene er man også i samsvar med forordningen.

6.4 Andre retningslinjer og veiledninger relevante for forsvarlig bruk av KI

Det finnes en rekke retningslinjer og veiledninger som er utviklet for å ivareta sikkerheten ved bruk av KI. Nedenfor følger en beskrivelse av de mest relevante.

- AMLAS (Assurance of Machine Learning for use in Autonomous Systems) /58/ er en metodikk utviklet for å integrere sikkerhetssikring i utviklingen av maskinlæringskomponenter. AMLAS dekker flere faser, inkludert sikkerhetssikring, datastyring og modellverifisering, og bidrar til å etablere et evidensgrunnlag for sikkerheten til KI-komponenter når de integreres i autonome systemer.
- DNV-RP-0671 «Assurance of AI-enabled systems» /19/ er en veiledning utarbeidet av DNV som spesifikt fokuserer på evaluering av KI-systemer.
- Studiene «Management System Support for Trustworthy Artificial Intelligence» /60/ og «Safe AI-how is this possible» /64 fra forskningsinstituttet Fraunhofer i Tyskland gir retningslinjer for implementering av pålitelige KI-systemer. De legger vekt på åpenhet, verifikasjon, validering og behovet for menneskelig tilsyn i kritiske KI-applikasjoner.
- NIST (National Institute of Standards and Technology) har utviklet retningslinjer for håndtering av bias i KI-systemer /61/. Disse retningslinjene fokuserer på å identifisere og begrense skjevheter i KI-algoritmer og data, noe som er avgjørende for å sikre rettferdighet og pålitelighet i sikkerhetskritiske anvendelser.
- TÜV i Tyskland publiserte «Trusted artificial intelligence towards certification of machine learning applications» /63/ som fokuserer på verifisering og validering av KI-systemer. Den vektlegger åpenhet, forklarbarhet og uavhengig tredjepartsvurdering for å sikre at KI-systemene fungerer som tiltenkt under forskjellige operasjonelle scenarier.

7 VIDERE ARBEIDE

7.1 Felles veiledninger for petroleumsindustrien

Havtils regelverk henviser til en lang rekke standarder og veiledninger, både Norske og internasjonale. Disse velger de fleste aktørene å følge, siden man ellers må demonstrere at alternative tilnærminger er like bra eller bedre. Bruk av felles standarder og veiledninger bidrar til et harmonisert sikkerhetsnivå i petroleumsindustrien. Som beskrevet i 2.2 og 5.3, finnes det imidlertid så langt ingen standarder eller veiledninger som er utarbeidet spesifikt med tanke på sikker bruk av KI innenfor petroleumsnæringen. For å opprettholde et harmonisert sikkerhetsnivå, og for å redusere bevisbyrden for den enkelte aktør, vil det være ønskelig at relevante aktører i industrien går sammen om å utarbeide veiledninger som representerer beste praksis for bruk av forskjellige typer av løsninger som inneholder KI.

Et slikt arbeide bør også kunne gjøre det enklere å møte kravene i EU forordningen om KI. Foreløpig må den enkelte aktør som ønsker å ta i bruk KI selv konkretisere og svare ut kravene i forordningen i egne styringssystemer. Slik operasjonalisering av høynivå krav er vanligvis arbeidskrevende, men arbeidsbyrden på hver enkelt organisasjon bør kunne reduseres dersom industrien går sammen om det.

7.2 Automatisert deteksjon av utrygge tilstander

Et gjentagende tema i denne rapporten, er at økt automatisering kan gjøre det vanskelig for operatører å forstå når det er nødvendig å aktivere barrierer manuelt. Innføring av KI forventes å kunne forsterke denne problematikken. Utfordringen knyttet til menneskelig deteksjon av utrygge tilstander, tilsier at industrien bør utforske mulighetene for automatisert deteksjon av utrygge tilstander forårsaket av KI.

7.3 Kvalifisering og vedlikehold av KI baserte systemer

For noen KI-systemer så vil resultatene som blir produsert ikke være deterministiske. Det betyr at dersom man utfører flere tester med samme inngangsdata, så får man ikke nødvendigvis de samme resultatene, noe som kan vanskeliggjøre kvalifisering, validering og også vedlikehold av programvare som inneholder KI.

Utfordringen knyttet til vedlikehold handler om at re-kjøring av tester for å sjekke at man får samme resultater som før, såkalt regresjonstesting, er den vanligste måte å sjekke at man ikke har fått uønskede side-effekter i forbindelse med modifikasjoner av programvare.

Dette kan også være kan være en begrensende faktor for hvor KI kan innføres, og industrien bør utforske hvordan denne utfordringen kan løses.

8 REFERANSER

/1/	Framework HSE Regulations, 18.12.2023, rammeforskriften_n.pdf (havtil.no).
/2/	Management Regulations, 18.12.2023, styringsforskriften_n.pdf (havtil.no).
/3/	Activities Regulations, 18.12.2023, aktivitetsforskriften_n.pdf (havtil.no).
/4/	Facilities Regulations, 18.12.2023, innretningsforskriften_n.pdf (havtil.no).
/5/	Technical and Operational Regulations, 18.12.2023, teknisk_og_operasjonell_forskrift_n.pdf (havtil.no).
/6/	ICT Security – Robustness in the Petroleum Sector – Regulations and Supervisory Methodology, DNV GL 2019-082, 24.02.2020.
/7/	Principles for barrier management in the petroleum industry – Barrier memorandum, Petroleum Safety Authority Norway 2017.
/8/	Integrated and unified risk management in the petroleum industry, Petroleum Safety Authority Norway, June 2018.
/9/	Oil & Gas UK Guidelines on Qualification of Materials for the Abandonment of Wells, issue 2.
/10/	NOG 070, Guidelines for the Application of IEC 61508 and IEC 61511 in the Norwegian Petroleum Industry (recommended SIL requirements). Guideline, Norwegian Oil and Gas Association, 2020.
/11/	NORSOK-I-002, Industrial automation and control systems, 2021.
/12/	DNV-RP-A203, Technology Qualification.
/13/	DNV-RP-A204, Assurance of digital twins.
/14/	DNV-RP-0317, Assurance of data collection and transmission in sensor systems.
/15/	DNV-RP-0497, Assurance of data quality management.
/16/	DNV-RP-0510, Assurance of data driven applications.
/17/	DNV-RP-0513, Assurance of simulation models.
/18/	DNV-RP-0665, Assurance of machine learning applications.
/19/	DNV-RP-0671, Assurance of AI-enabled systems.
/20/	D. Gupta, M. Shah (2022). A comprehensive study on artificial intelligence in the oil and gas sector. Environ Sci Pollut Res Int. 2022 Jul; 29(34):50984-50997. doi: 10.1007/s11356-021-15379-z (Retracted).
/21/	N. Aissani, B. Beldjilali, D. Trentesaux (2009), Dynamic scheduling of maintenance tasks in the petroleum industry: a reinforcement approach. Eng Appl Artif Intell 22(7):1089–1103. https://doi.org/10.1016/j.engappai.2009.01.014.

/22/	M. Alhashem (2019), Supervised machine learning in predicting multiphase flow regimes in horizontal pipes. Society of Petroleum Engineers - Abu Dhabi International Petroleum Exhibition and Conference 2019. ADIP 2019. https://doi.org/10.2118/197545-ms.
/23/	Bello O, Teodoriu C, Yaqoob T, Oppelt J, Holzmann J, Obiwanne A (2016). Application of artificial intelligence techniques in drilling system design and operations: a state of the art review and future research pathways. Society of Petroleum Engineers - SPE Nigeria. Annual International Conference and Exhibition. https://doi.org/10.2118/184320-ms.
/24/	R. Brelsford (2018). Repsol launches big data, AI project at Tarragona refinery. https://www.ogj.com/refining-processing/refining/operations/article/17296578/repsol-launches-big-data-ai-project-attarragona-refinery.
/25/	Luca Cadei, Marco Montini, Fabio Landi, Francesco Porcelli, Vincenzo Michetti, Matteo Origgi, Marco Tonegutti, Sylvain Duranton (2019). Big data advanced analytics to forecast operational upsets in upstream production system. Society of Petroleum Engineers - Abu Dhabi International Petroleum Exhibition and Conference 2018. ADIPEC 2018. https://doi.org/10.2118/193190-ms.
/26/	S. Chaki, A.K. Verma, A. Routray, W.K. Mohanty, M. Jenamani (2014). Well tops guided prediction of reservoir properties using modular neural network concept: a case study from western onshore, India. J Pet Sci Eng 123:155–163. https://doi.org/10.1016/j.petrol.2014.06.019.
/27/	V.L.C. de Oliveira, A.P.M. Tanajura, H.A. Lepikson (2013). A multi-agent system for oil field management. 11th IFAC Workshop on Intelligent Manufacturing Systems. The International Federation of Automatic Control. 1-6.
/28/	Y. Gidh, A. Purwanto, H. Ibrahim (2012). Artificial neural network drilling parameter optimization system improves ROP by predicting/managing bit wear. Society of Petroleum Engineers – SPE Intelligent Energy International 2012(1):195–207. https://doi.org/10.2118/149801-ms.
/29/	F.A.S. Adesina, A. Abiodun, A. Anthony, F. Olugbenga (2015). Modelling the effect of temperature on environmentally safe oil based drilling mud using artificial neural network algorithm. Petroleum and Coal Journal, Volume 57, Number 1, pp. 60-70.
/30/	A. Ahmed, S. Elkatatny, A. Ali (2021). Fracture pressure prediction using surface drilling parameters by artificial intelligence techniques. J Energy ResourceTechnolgy 143(3):20.
/31/	V. Baskaran, S. Singh, V. Reddy, S. Mohandas (2019). Digital assurance for oil and gas 4.0: role, implementation and case studies. In SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition, p 20. https://www.scilit.net/publications/17a663acdcaf179b3a98af9ce989be51.
/32/	S. Choubey, G.P Karmakar (2021). Artificial intelligence techniques and their application in oil and gas industry. Artif. Intell. Rev. 54(5):3665-3683.
/33/	Christopher Jeffery and Andrew Creegan (2020). Adaptive drilling application uses AI to enhance on-bottom drilling performance.
/34/	Yue Wang, Sai Ho Chung (2021), Artificial intelligence in safety-critical systems: a systematic review.
/35/	W. Khalid, I. Soleymani, N.H. Mortensen, K.V. Sigsgaard (2021). AI-based maintenance scheduling for offshore oil and gas platforms. Annual Reliability and Maintainability Symposium (RAMS), p. 1-6.
/36/	L. Kirschbaum, D. Roman, G. Singh, J. Bruns, V. Robu, D. Flynn (2020). AI-driven maintenance support for downhole tools and electronics operated in dynamic drilling environments. IEEE Access, 8, 78683-78701.

/37/	European Commission (2019). Ethics guidelines for trustworthy AI. Technical report. European Commission's High-Level Expert Group on Artificial Intelligence.
/38/	Michael Bortz, Kai Dadhe, Sebastian Engell, Vanessa Gepert, Norbert Kockmann, Ralph Müller-Pfefferkorn, Thorsten Schindler, and Leon Urbas (2023). AI in Process Industries – Current Status and Future Prospects. Chem. Ing. Tech, 95, No. 7, 975-988.
/39/	S.W. Choi, E.B. Lee, J.H. Kim (2021). The engineering machine-learning automation platform (EMAP): a big-data-driven AI tool for contractors' sustainable management solutions for plant projects. Sustainability 13(18):365-384.
/40/	Alexander Inzartsev, Alexander Pavin, Alexander Kleschev, Valeria Gribova (2016). Application of artificial intelligence techniques for fault diagnostics of autonomous underwater vehicles.
/41/	S. Heshmati-alamdari, A. Eqtami, G.C. Karras, D.V. Dimarogonas, K.J. Kyriakopoulos (2020). A self-triggered position based visual serving model predictive control scheme for underwater robotic vehicles. Machines, 8, 33.
/42/	Daniel Mitchell, Jamie Blanche, Sam Harper, Theodore Lim, Ranjeetkumar Gupta, Osama Zaki, Wenshuo Tang, Valentin Robu, Simon Watson, David Flynn (2022). A review: Challenges and opportunities for artificial intelligence and robotics in the offshore wind sector
/43/	Alf Inge Molde (2018). Intelligent progress. Norwegian Continental Shelf 1-2018, p32-36.
/44/	Håvard Devold, Roar Fjellheim (2019). Artificial intelligence in autonomous operation of oil and gas facilities. Society of Petroleum Engineers. SPE-197399-MS.
/45/	Dmitry Koroteev, Zeljko Tekic (2021). Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future. https://doi.org/10.1016/j.egyai.2020.100041. Energy and AI, Volume 3.
/46/	S. Bernardini, F. Jovan, Z. Jiang, P. Moradi, T. Richardson, R, Sadeghian, S. Sareh, S. Watson, A. Weightman (2020). A multi-robot platform for the autonomous operation and maintenance of offshore wind farms. In Autonomous Agents and Multi-Agent Systems (AAMAS) 2020 International Foundation for Autonomous Agents and Multiagent Systems.
/47/	Håvard Devold, Roar Fjellheim (2019). Artificial Intelligence in Autonomous Operation of Oil and Gas Facilities. Society of Petroleum Engineers, SPE-197399-MS.
/48/	Y.K. Dwivedi, L. Hughes, E. Ismagilova, et al. (2021). Artificial Intelligence (AI): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice, and policy. International Journal of Information Management.
/49/	Ministry of Local Government and Modernisation (2020). National strategy for artificial intelligence.
/50/	EU Artificial Intelligence Act, COM/2024/1689, http://data.europa.eu/eli/reg/2024/1689/oj.
/51/	NEK ISO/IEC 5338:2023. Information technology – Artificial intelligence – AI system life cycle processes.
/52/	NEK ISO/IEC TR 5469:2024. Artificial intelligence – Functional safety and AI systems.
/53/	NS-EN ISO/IEC 22989:2023. Information technology – Artificial intelligence – Concepts and terminology for artificial intelligence.
/54/	NEK ISO/IEC 23894:2023. Information technology – Artificial intelligence – Guidance on risk management.

/55/	NEK ISO/IEC 42001:2023. Information technology – Artificial intelligence – Management system.
/56/	NEK IEC 61508:2010. Functional safety of electrical/electronic/programmable electronic safety-related systems.
/57/	NEK IEC 61511:2016. Functional safety – Safety instrumented systems for the process industry sector.
/58/	Richard Hawkins, Colin Paterson, Chiara Picardi, Yan Jia, Radu Calinescu and Ibrahim Habli (2021). Guidance on the assurance of machine learning in autonomous systems (AMLAS). Assuring Autonomy International Programme (AAIP). University of York.
/59/	University of Oxford (2022). capAI; A procedure for conducting conformity assessment of AI systems in line with the EU Artificial Intelligence Act.
/60/	Fraunhofer IAIS (2021). Management system support for trustworthy artificial intelligence; A comparative study.
/61/	NIST (2022). Towards a standard for identifying and managing bias in artificial intelligence.
/62/	Consortium of Quality Assurance (2021). Guidelines for quality assurance of AI-based products and services.
/63/	TÜV (2021). Trusted artificial intelligence towards certification of machine learning applications.
/64/	Dr. Harald Rueß, Prof. Dr. Simon Burton (2022). Fraunhofer IKS. Safe AI—How is this possible?
/65/	J. Smith, L. Brown (2023). Artificial Intelligence in Oil and Gas: Safety and Surveillance Applications. Journal of Petroleum Technology, 2023.
/66/	Kizzy Nkem Elliot, Levi Adawari Damingo (2024), Application of artificial intelligence in the oil and gas industry. International Research Journal of Modernization in Engineering Technology and Science. Volume 6, issue 5.
/67/	Y. Zhang, H. Lee (2022). Real-time safety monitoring in offshore wind farms using AI. Renewable Energy.
/68/	Masoud Masoumi (2023). Machine Learning Solutions for Offshore Wind Farms: A Review of Applications and Impacts. J. Mar. Sci. Eng. 2023, 11(10), 1855; https://doi.org/10.3390/jmse11101855.
/69/	Directive 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery.
/70/	Directive 2014/34/EU of the European Parliament and of the Council of 26 February 2014 on the harmonisation of the laws of the Member States relating to equipment and protective systems intended for use in potentially explosive atmospheres.
/71/	Training AI requires more data than we have — generating synthetic data could help solve this challenge (theconversation.com).
/72/	Christoffersen, K., & Woods, D. D. (2002). How to make automated systems team players. In Advances in Human Performance and Cognitive Engineering Research (Vol. 2, pp. 1–12). Emerald Group Publishing Limited. https://doi.org/10.1016/S1479-3601(02)02003-9.
/73/	Littman, M., Ajunwa, I., Berger, G., Boutilier, C., Currie, M., Doshi Velez, F., Hadfield, G., Horowitz, M., Isbell, C., Kitano, H., Levy, K., Lyons, T., Mitchell, M., Shah, J., Sloman, S., Vallor, S., & Walsh, T. (2021). Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report. Stanford University. http://ai100.stanford.edu/2021-report.

/74/	National Academies of Sciences, Engineering and Medicine. (2022). Human-AI Teaming: State of the Art and Research Needs. The National Academies Press. https://doi.org/10.17226/26355.
/75/	Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned from Human-Automation Research. Human Factors, 59(1), 5–27. https://doi.org/10.1177/0018720816681350.
/76/	Hollnagel, E. (2012). Coping with complexity: Past, present and future. Cognition, Technology & Work, 14(3), 199–205. https://doi.org/10.1007/s10111-011-0202-7.
/77/	Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8.
/78/	Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. Human Factors, 37(2), 381–394. https://doi.org/10.1518/001872095779064555.
/79/	Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. Human Factors: The Journal of the Human Factors and Ergonomics Society, 52(3), 381–410. https://doi.org/10.1177/0018720810376055.
/80/	Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and Automation Bias in the Use of Imperfect Automation. Human Factors: The Journal of the Human Factors and Ergonomics Society, 57(5), 728–739. https://doi.org/10.1177/0018720815581940.
/81/	Mosier, K., & Skitka, L. (1996). Human Decision Makers and Automated Decision Aids: Made for Each Other? In R. Parasuraman & M. Mouloua (Eds.), Automation and human performance: Theory and applications (Vol. 40, pp. 201–220). Lawrence Erlbaum Associates, Inc.
/82/	Finomore, V. S., Shaw, T. H., Warm, J. S., Matthews, G., & Boles, D. B. (2013). Viewing the workload of vigilance through the lenses of the NASA-TLX and the MRQ. Human Factors, 55(6), 1044–1063. https://doi.org/10.1177/0018720813484498.
/83/	Warm, J. S., Dember, W. N., & Hancock, P. A. (1996). Vigilance and workload in automated systems. In Automation and human performance: Theory and applications. (pp. 183–200). Lawrence Erlbaum Associates, Inc.
/84/	Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. Human Factors, 39(2), 230–253. https://doi.org/10.1518/001872097778543886.
/85/	Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. Human Factors: The Journal of the Human Factors and Ergonomics Society, 56(3), 476–488. https://doi.org/10.1177/0018720813501549.
/86/	Endsley, M. R. (2023). Ironies of artificial intelligence. Ergonomics, 0(0), 1–13. https://doi.org/10.1080/00140139.2023.2243404.
/87/	Endsley, M. R. (2023). Supporting Human-AI Teams: Transparency, explainability, and situation awareness. Computers in Human Behavior, 140, 107574. https://doi.org/10.1016/j.chb.2022.107574.
/88/	Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392.

/89/	Bach, T. A., Kristiansen, J. K., Babic, A., & Jacovi, A. (2024). Unpacking Human-AI Interaction in Safety-Critical Industries: A Systematic Literature Review. IEEE Access, 12, 106385–106414. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3437190.
/90/	Kim, J. W., & Jung, W. (2003). A taxonomy of performance influencing factors for human reliability analysis of emergency tasks. Journal of Loss Prevention in the Process Industries, 16(6), 479–495. https://doi.org/10.1016/S0950-4230(03)00075-5.
/91/	Rouse, W. B., & Morris, N. M. (1986). On looking into the black box: Prospects and limits in the search for mental models. Psychological Bulletin, 100(3), 349–363. https://doi.org/10.1037/0033-2909.100.3.349.
/92/	Greenlee, E. T., DeLucia, P. R., & Newton, D. C. (2022). Driver Vigilance Decrement is More Severe During Automated Driving than Manual Driving. Human Factors. https://doi.org/10.1177/00187208221103922.
/93/	Naujoks, F., Purucker, C., Wiedemann, K., & Marberger, C. (2019). Noncritical State Transitions During Conditionally Automated Driving on German Freeways: Effects of Non–Driving Related Tasks on Takeover Time and Takeover Quality. Human Factors, 61(4), 596–613. https://doi.org/10.1177/0018720818824002.
/94/	Veitch, E., Christensen, K., Log, M., Valestrand, E., Hilmo Lundheim, S., Nesse, M., Alsos, O., & Steinert, M. (2022). From captain to button-presser: Operators' perspectives on navigating highly automated ferries. Journal of Physics: Conference Series, 2311, 012028. https://doi.org/10.1088/1742-6596/2311/1/012028.
/95/	Endsley, M. R., Bolté, B., & Jones, D. G. (2003). Designing for situation awareness: An approach to user-centered design. Taylor & Francis.
/96/	Sheridan, T. B. (2021). Human Supervisory Control of Automation. In G. Salvendy & W. Karwowski (Eds.), Handbook of Human Factors and Ergonomics (5th ed., pp. 736–760). Wiley. https://doi.org/10.1002/9781119636113.ch28.
/97/	Wickens, C. D., & Carswell, C. M. (2021). Information processing. In G. Salvendy & W. Karwowski (Eds.), Handbook of Human Factors and Ergonomics (5th ed., p. 1603). John Wiley & Sons, Incorporated.
/98/	ISO. (2019). ISO 9241-210:2019 Ergonomics of human-system interaction—Part 210: Human-centred design for interactive systems. International Organization for Standardization.
/99/	Hoem, Å. S., Veitch, E., & Vasstein, K. (2022). Human-centred risk assessment for a land-based control interface for an autonomous vessel. WMU Journal of Maritime Affairs, 21(2), 179–211. https://doi.org/10.1007/s13437-022-00278-y.
/100/	OROK. (2018). S-002:2018+AC:2021 Working environment. Standard Norway.
/101/	van de Merwe, K., Mallam, S., Engelhardtsen, Ø., & Nazir, S. (2023). Towards an approach to define transparency requirements for maritime collision avoidance. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 67(1), 483–488. https://doi.org/10.1177/21695067231192862.
/102/	van de Merwe, K., Mallam, S., Nazir, S., & Engelhardtsen, Ø. (2024). The Influence of Agent Transparency and Complexity on Situation Awareness, Mental Workload, and Task Performance. Journal of Cognitive Engineering and Decision Making, 18(2), 156–184. https://doi.org/10.1177/15553434241240553.
/103/	Norman, D. A. (1990). The "problem" with automation: Inappropriate feedback and interaction, not "over-automation." Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 327(1241), 585–593. https://doi.org/10.1098/rstb.1990.0101.

/104/	van Doorn, E., Horváth, I., & Rusák, Z. (2021). Effects of coherent, integrated, and context-dependent adaptable user interfaces on operators' situation awareness, performance, and workload. Cognition, Technology & Work, 23(3), 403–418. https://doi.org/10.1007/s10111-020-00642-z.
/105/	Bach, T. A., Kristiansen, J. K., Babic, A., & Jacovi, A. (2023). Unpacking Human-AI Interaction in Safety-Critical Industries: A Systematic Literature Review (arXiv:2310.03392). arXiv. http://arxiv.org/abs/2310.03392.
/106/	Beck, H. P., Dzindolet, M. T., & Pierce, L. G. (2007). Automation Usage Decisions: Controlling Intent and Appraisal Errors in a Target Detection Task. Human Factors, 49(3), 429–437. https://doi.org/10.1518/001872007X200076.
/107/	Ezenyilimba, A., Wong, M., Hehr, A., Demir, M., Wolff, A., Chiou, E., & Cooke, N. (2023). Impact of Transparency and Explanations on Trust and Situation Awareness in Human–Robot Teams. Journal of Cognitive Engineering and Decision Making, 17(1), 75–93. https://doi.org/10.1177/15553434221136358.
/108/	Oliveira, L., Burns, C., Luton, J., Iyer, S., & Birrell, S. (2020). The influence of system transparency on trust: Evaluating interfaces in a highly automated vehicle. Transportation Research Part F: Traffic Psychology and Behaviour, 72, 280–296. https://doi.org/10.1016/j.trf.2020.06.001.
/109/	Warden, T., Carayon, P., Roth, E. M., Chen, J., Clancey, W. J., Hoffman, R., & Steinberg, M. L. (2019). The National Academies Board on Human System Integration (BOHSI) Panel: Explainable AI, System Transparency, and Human Machine Teaming. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63(1), 631–635. https://doi.org/10.1177/1071181319631100.
/110/	van de Merwe, K., Mallam, S., & Nazir, S. (2024). Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review. Human Factors, 66(1), 180–208. https://doi.org/10.1177/00187208221077804.
/111/	Strauch, B. (2018). Ironies of Automation: Still Unresolved After All These Years. IEEE Transactions on Human-Machine Systems, 48(5), 419–433. https://doi.org/10.1109/THMS.2017.2732506

APPENDIKS A. EKSEMPEL: KI-DREVET PREDIKTIVT VEDLIKEHOLD FOR OFFSHORE BOREPLATTFORMER

Bakgrunn

Offshore boreoperasjoner er sentrale for energisektoren, men har med komplekse utfordringer og iboende risikoer på grunn av det operasjonelle miljøet, den mekaniske kompleksiteten til boreutstyret, og de høye kostnadene forbundet med utstyrsfeil og nedetid. I denne sammenhengen fremstår prediktivt vedlikehold som en relevant strategi for å forutse og forhindre feil før de oppstår.

KI-implementering

Implementering av et KI-drevet system for prediktivt vedlikehold på offshore boreplattformer innebærer utplassering av et nettverk av sensorer og IoT-enheter for å kontinuerlig overvåke tilstanden til kritiske utstyrskomponenter i sanntid. Disse dataene inkluderer vibrasjon, temperatur, trykk, og andre operasjonelle parametere som er kritiske for boreutstyrets funksjon.

Maskinlæringsalgoritmer analyserer disse dataene, lærer av historiske vedlikeholdsrapporter, feilforekomster, og operasjonelle anomalier for å forutse potensielle feil eller behov for vedlikehold før en kritisk feil oppstår. Disse algoritmene kan ikke bare oppdage mønstre som indikerer fremvoksende feil, men kan også foreslå optimale vedlikeholdsskjemaer, dermed reduseres unødvendige inspeksjoner og reparasjoner, og driftsavbrudd minimeres.

Fordeler

1. Ved å forutse utstyrsfeil før de inntreffer, reduseres risikoen for ulykker og farlige forhold for operatørene
2. Nedetid på grunn av uplanlagt vedlikehold og reparasjoner minimeres, noe som sikrer at boreoperasjonene kjører jevnere og mer effektivt.
3. KI-drevet prediktivt vedlikehold kan føre til betydelige kostnadsbesparelser gjennom å redusere behovet for korrektivt vedlikehold og forlenge levetiden til kritisk boreutstyr.

Utfordringer

- Sømløs integrering av KI-systemer med eksisterende offshore-infrastruktur.
- Å sikre pålitelighet og nøyaktighet av KI-prediksjoner for å unngå falske positiver eller negativer. Dette vil være særlig kritisk dersom det KI drevne systemet også bestemmer vedlikeholdsintervallene.
- Å adressere cybersikkerhetsbekymringer knyttet til implementering av IoT og KI.
- Opplæring av personell til å samhandle med og reagere på KI-drevne vedlikeholds anbefalinger.
- Å sikre at KI-implementeringer følger eksisterende reguleringer og standarder for sikkerhetskritiske systemer i offshoreindustrien.

Om DNV

Vi er et globalt selskap innen kvalitetssikring og risikohåndtering med tilstedeværelse i over 100 land. Vårt formål er å sikre liv, verdier og miljøet. Med vår unike tekniske ekspertise og uavhengighet bistår vi våre kunder med å forbedre sikkerhet, effektivitet og bærekraft.

Enten vi godkjenner et nytt skipsdesign, optimerer energiproduksjonen fra en vindmøllepark, analyserer sensordata fra en gassrørledning eller sertifiserer verdikjeden til en matprodusent, hjelper vi våre kunder med å ta gode og riktige beslutninger og øke tilliten til virksomheten, produktene og tjenestene deres. Verden er i endring. Vi kan påvirke utviklingen. Sammen skal vi takle de globale utfordringene og omstillingene vi vil møte.