Håndtering af bias i kunstig intelligens
- en introduktion til DS/PAS 2500-3:2023

Program

13.00-13.20

Velkomst og introduktion til specifikationen om bias.
DS/PAS 2500-3: 2023, Kunstig intelligens
– Del 3: Bias

13.20-13.50

Sådan håndterer du bias i kunstig intelligens – paneldebat

13.50-14.00

Q&A – få svar på dine spørgsmål

14.00

Tak for i dag

Ekspert panel

Marie Valentin Beck

Diversitetsstrategisk rådgiver

Bureau M

Thomas Hildebrandt

Professor

DIKU

Anders Kofod-Petersen

Professor (NTNU)

Piedboeuf

En verden uden standarder

Billeder fra Standard Norge.

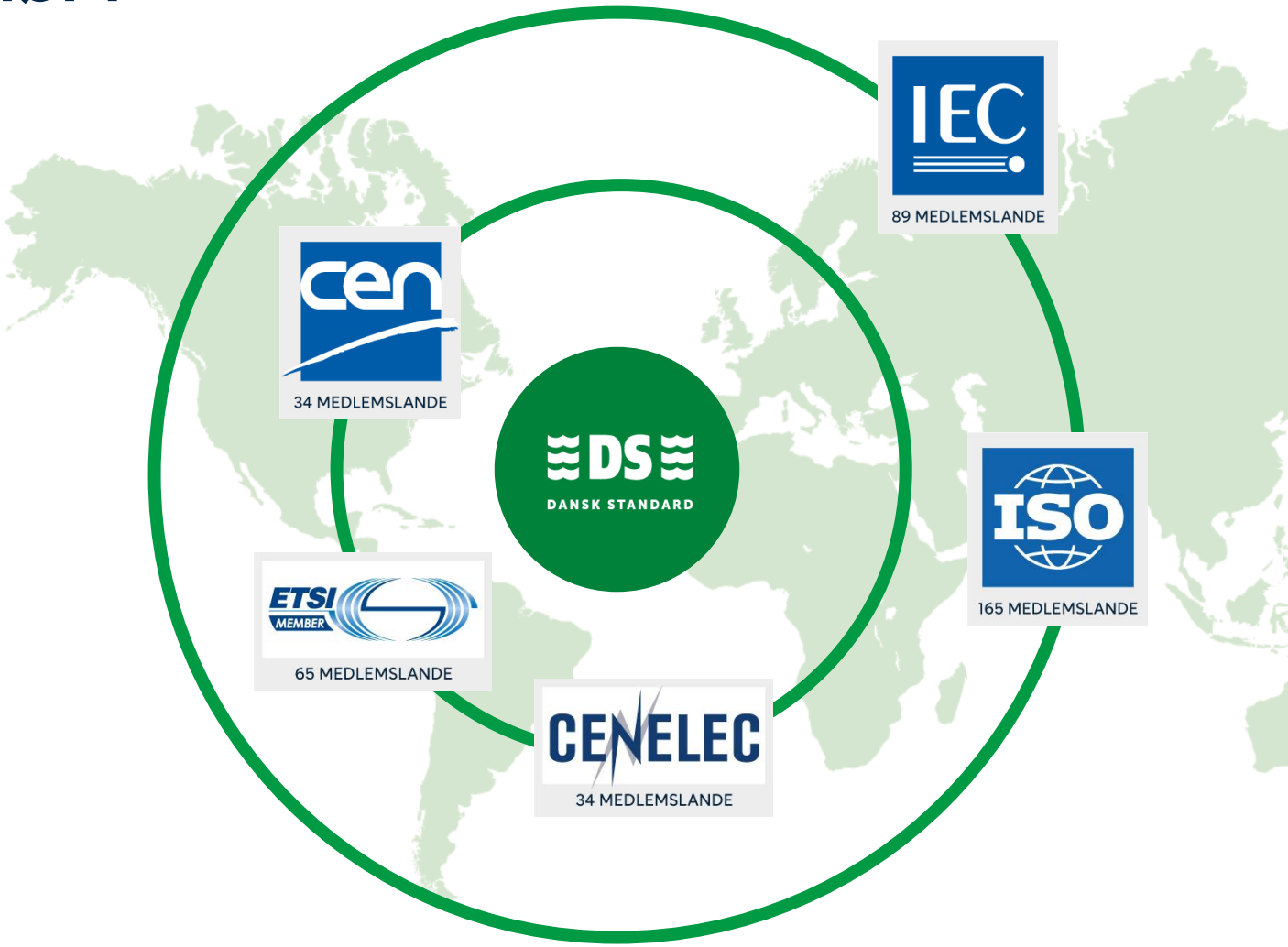
Standarder skaber tryghed i samfundet

- Standarder skaber **sikre produkter**, som virksomheder, borgere og myndigheder kan have **tillid til og bruge**
- gennem en **entydig fælles forståelse** af **krav, sikkerhed og kvalitet**
- der letter din hverdag.

Dansk Standard – en stærk aktør i et europæisk og globalt standardiseringsnetværk

ISO har 164 medlemslande som medlem ud af 194 lande i verden.

Mere end 100.000 deltagere i mere end 3.000 arbejdsgrupper.



National AI strategi

…Regeringen vil igangsætte et arbejde med at ***udvikle nationale tekniske specifikationer,*** *som tager udgangspunkt i danske virksomheders konkrete behov. Arbejdet vil blandt andet tage udgangspunkt i de seks etiske principper for kunstig intelligens.*

Regeringen, marts 2019

DS/PAS - en serie af specifikationer om kunstig intelligens

Hvad er en DS/PAS?

- PAS er en forkortelse for 'Publicly Available Specification' som er en publikation, der er udarbejdet på nationalt niveau og ikke har status som en international standard.
- En PAS adskiller sig fra en international standard ved f.eks. ikke at have de samme krav til graden af interessent-involvering eller opsætning.
- En DS/PAS specificerer ikke krav, der skal overholdes, men indeholder anbefalinger, information og gode råd.

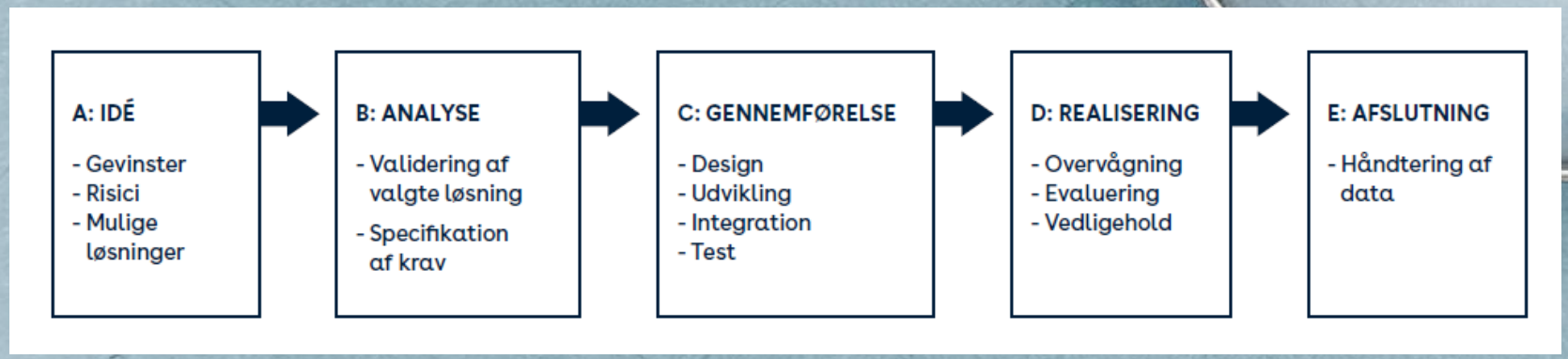
Processen

Milepæle april 2022 - april 2023

- To workshops afholdt - 60+ eksperter deltog
- Kommenteringsfase med 100 interessenter
- Interessenter afgav mere end 100 kommentarer
- Overordnet feedback var positiv og meget konstruktiv
- Publicering i april 2023 via ds.dk/ai og webshop.ds.dk

Specifikationens opbygning (baseret på ISO/CEN skabelon)

1. Anvendelsesområde
2. Normative reference
3. Termer og definitioner
4. Bias
   - Menneskelig bias
   - Databias
5. Tjeklister for hver fase i et typisk IT projekt



1. Anvendelsesområde
2. Normative reference
3. Termer og definitioner
4. Bias
   - Menneskelig bias
   - Databias
5. Tjeklister for hver fase i et typisk IT projekt

DANSK STANDARD

3 Termer og definitioner

I dette dokument gælder følgende termer og definitioner.

3.1
algoritme
beskrivelse af en proces til at løse et generelt problem

3.2
kunstig intelligens
systems evne til at tilegne, behandle og anvende *viden* (3.3), og/eller færdigheder til at finde løsninger på problemer, træffe eller støtte beslutninger

Note 1 til term: Kunstig intelligens sammenlignes som regel med menneskelig intelligens, hvilket dog ikke er udtømmende for begrebet.

Note 2 til term: Maskinlæring er en underkategori af kunstig intelligens.

3.3
viden
data, information og færdigheder tilegnet via erfaring eller læring

Definitioner af bias

3.7
bias
systematisk forskel eller fejl i behandling af data

3.14
databias
dataegenskaber, der, hvis de ikke adresseres, kan føre til et system, der har *bias* (3.7) i forhold til forskellige grupper

3.17
menneskelig bias
bias (3.7), der opstår, når mennesker behandler og fortolker information

1. Anvendelsesområde
2. Normative reference
3. Termer og definitioner
4. Bias
   - Menneskelig bias
   - Databias
5. Tjeklister for hver fase i et typisk IT projekt

DANSK STANDARD

4 Bias

4.1 Introduktion til bias

Denne specifikation forstår bias som systematisk forskel i behandling af data, som kan have den effekt, at der sker en forskel i behandlingen af specifikke objekter, personer eller grupper i forhold til andre. Det kan fx være mellem personer af forskellige køn og alder eller mellem personer med forskellig herkomst. En behandling kan være enhver form for handling, herunder opfattelse, observation, klassifikation, repræsentation, forudsigelse eller beslutning. Enhver form for ikke-triviel behandling vil indeholde bias, da behandlingen ellers vil være uafhængig af det objekt, der behandles. Det er derfor relevant at tale om, hvilken effekt den pågældende bias har, og om denne effekt er positiv (ønsket), om den er negativ (uønsket) eller om den er neutral.

1. Anvendelsesområde
2. Normative reference
3. Termer og definitioner
4. Bias
   - Menneskelig bias
   - Databias
5. Tjeklister for hver fase i et typisk IT projekt

Tjeklister for hver fase

DANSK STANDARD

5.4.2 Fase A: Idéfasen

Organisationen med ansvar for idéfasen bør som minimum overveje og dokumentere svarene på spørgsmålene i A1-A4 for hver tænkt anvendelse, med inddragelse af relevante interessenter eller repræsentanter for disse.

ID	SPØRGSMÅL	HANDLING
A1	**Hvilke grupper i beslutningsdomænet for systemet er relevante i forhold til bias?**	[Anvend gerne risicimodel og/eller PDCA-metoden beskrevet i anneks C og D]
	Svar (udfyld felter)	**Handling** (udfyld felter)
A1-1: Projektejer/ kravstiller		
A1-2: Slutbrugere		
A1-3: Borgere (både indirekte og direkte berørte)		
A1-4: Leverandører		
A1-5: Udviklere		
A1-6: Retsligt personale/tilsynsførende/uvildige tekniske eksperter		

C.1 Risicimodel lavet på baggrund af DS/EN IEC 31010:2019, *Risikoledelse – Teknikker til risikovurdering*

Følgende model kan anvendes til at vurdere risikoen i de enkelte faser i projektet.

Risiko	**Kort beskrivelse af risici**	**Konsekvens** (1 = lille; 3 = mellem; 5 = høj)	**Sandsynlighed** (1 = lille; 3 = mellem; 5 = høj)	**Risikofaktor** (konsekvens x sandsynlighed)	**Hvordan kan vi mitigere?**
#1					
#2					
#3					

Tabel C.1 - Risikoskema til udfyldelse

DS
DANSK STANDARD

Q&A
– stil dit spørgsmål i chatten

Få svar på dine spørgsmål om specifikationen og det øvrige standardiseringsarbejde.

DANSK STANDARD

Opsamling

Vil du vide mere

Kontakt:

Kim Skov Hilding
Seniorkonsulent
Mail: ksk@ds.dk

Ditte Heede
Konsulent
Mail: dkh@ds.dk

Tak for i dag