

Hvordan lykkes vi med ansvarlig KI?

Retningslinjers rolle i å utvikle et felles språk og praksis

Marianne J. Bjerkman, Høgskolen i Østfold



September 2027: Agent-4, the Superhuman AI Researcher

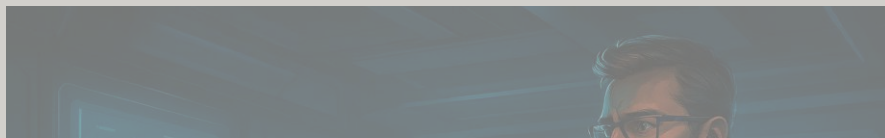
The gap between human and AI learning efficiency is rapidly decreasing.

Traditional LLM-based AIs seemed to require many orders of magnitude more data and compute to get to human level performance.⁷⁴ Agent-3, having excellent knowledge of both the human brain and modern AI, has created thousands of copies doing research, ends up making progress, narrowing the gap to an agent that's only around 4 times slower than the human brain.⁷⁵

This new AI system is dubbed Agent-4.

An individual copy of the model, running at human level, is much better at AI research than any human. 300,000 copies, running at the thinking speed of humans. Inside the corporation, from these copies, a year passes every week.⁷⁶ This system manages to speed up the overall rate of algorithmic progress. OpenBrain is heavily bottlenecked on compute to achieve a year's worth of algorithmic progress even when it's up against the limits of the Agent-4 paradigm.

As Agent-4 gets smarter, it becomes harder for Agent-3 to oversee it. For example, Agent-4's neuralese "language" becomes as alien and incomprehensible to Agent-3 as Agent-3's is to humans. Besides, Agent-4 is now much more capable than Agent-3 and has a good sense of exactly how to look good to it.



These researchers go to bed every night and wake up to another week worth of progress made mostly by the AIs. They work increasingly long hours and take shifts around the clock just to keep up with progress—the AIs never sleep or rest. They are burning themselves out, but they know that these are the last few months that their labor matters.

Bildet er generert av Copilot

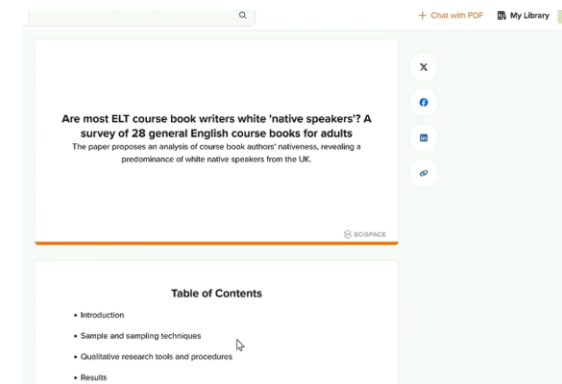
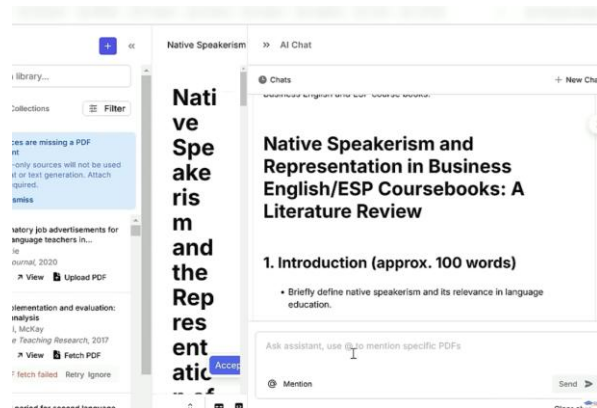
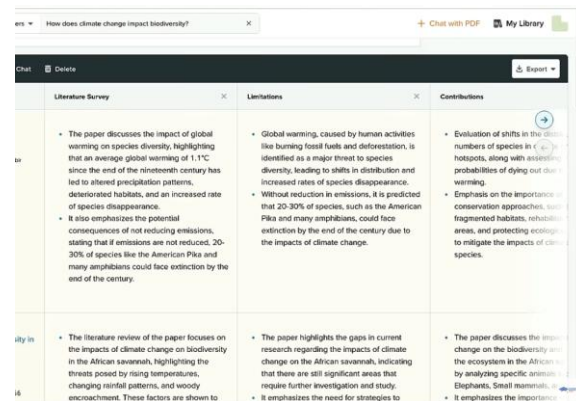
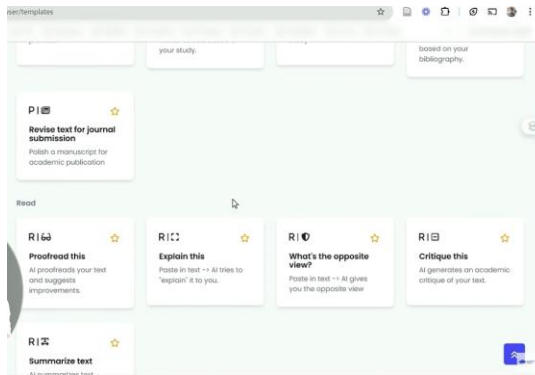
“

**Research is one of the
sectors that could be most
significantly disrupted by
generative AI (2)**

**Kan akselerere
vitenskapelige
oppdagelser**

**Kan endre måten
forskning utføres på**

**Kan innføre risiko for
påvirkning av
forskningsintegritet og
metodisk kvalitet**



Fremtidens

1 Norge –

I 2030 er Norge
den digitale or
innbyggere en
konkurransedy



gjennomført høsten
Regjeringen vil at 80
i løpet av 2025,
digitaliseringsstrategien.

nye og forbedre
kvaliteten på
trengs mest, sier
(Ap).

Bildet er generert av Copilot

AI's research blunder: How a mistake sparked a chain of flawed scientific papers. Can artificial intelligence be trusted in academia?

By Paurush Omar, ET Online • Last Updated: Mar 09, 2025, 08:22:00 PM IST

Synopsis

A viral video reveals how an AI-generated mistake led to nearly two dozen flawed research papers, sparking concerns about automation in academia. The error, originating from a mistranslation in a 1959 paper, went unnoticed by peer reviewers. This incident underscores the risks of unchecked AI in research and the necessity for rigorous human oversight.



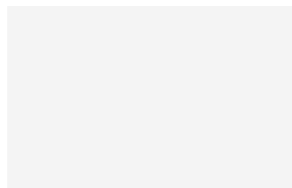
KUNSTIG INTELLIGENS

Ekspertene frykter at KI gjør studentene dummere

Flere studier tyder på at bruk av kunstig intelligens (KI) svekker studentenes læring. — For meg blir det dumt å ikke bruke det, sier student Francesca Savalli.



Kilder: (6), (7), (8), (9), (10)



KHrono

Nyheter Debatt Student Nytt om

KUNSTIG INTELLIGENS

I denne artikkelen ligger hemmelig beskjed. Slik skulle fagfellene lure

En avis fant skjulte instruksjoner i 17 forskningsartikler. Det er et syn at det globale forskningssystemet er ødelagt, mener informatiker Mo Irgens.



r/changemyview • for 2 md. siden
AutoModerator

META: Unauthorized Experiment on CMV Involving AI-generated Comments

META

The CMV Mod Team needs to inform the CMV community about an unauthorized experiment conducted by researchers from the University of Zurich on CMV users. This experiment deployed AI-generated comments to study how AI could be used to change views.

CMV rules do not allow the use of undisclosed AI generated content or bots on our sub. The researchers did not contact us ahead of the study and if they had, we would have declined. We have requested an apology from the researchers and asked that this research not be published, among other complaints. As discussed below, our concerns have not been substantively addressed by the University of Zurich or the researchers.

You have a right to know about this experiment. Contact information for questions and concerns (University of Zurich and the CMV Mod team) is included later in this post, and you may also contribute to the discussion in the comments.

KUNSTIG INTELLIGENS

ChatGPT formidlet falsk forskning

ChatGPT hadde ikke fått med seg at hodeskallen fra det som angivelig var verdens første gepard-art, var en forfalskning.



80%

...av studentene våre bruker KI!*

66%

...har ikke fått opplæring!*

September 2023; Nature*

Mai 2024; Oxford University Press*

Highlights

- Usage of generative AI tools is lower among women (7%), advanced researchers (19%), and individuals without specific training (8%).
- Encountering barriers is associated with an 11% increase in generative AI tool usage.
- Higher usage is observed among researchers in for-profit companies (19%), medical research (16%), and hospitals (15%).
- Government advisors utilize generative AI tools 45% more frequently than those in typical government roles.

Our key findings include:

1. Majority say they have used some form of AI

76% said they have used a form of AI tool in their research—with machine translation and chatbots being the most popular.

3. Vast majority say they are suspicious of AI companies

5. Fears about intellectual property

7. Generational differences in researchers' opinions of AI

9. Confusion around available guidance

2. Over 2 in 3 have felt the benefits of using AI

4. Concerns about AI affecting research quality

6. AI may reduce critical thinking skills

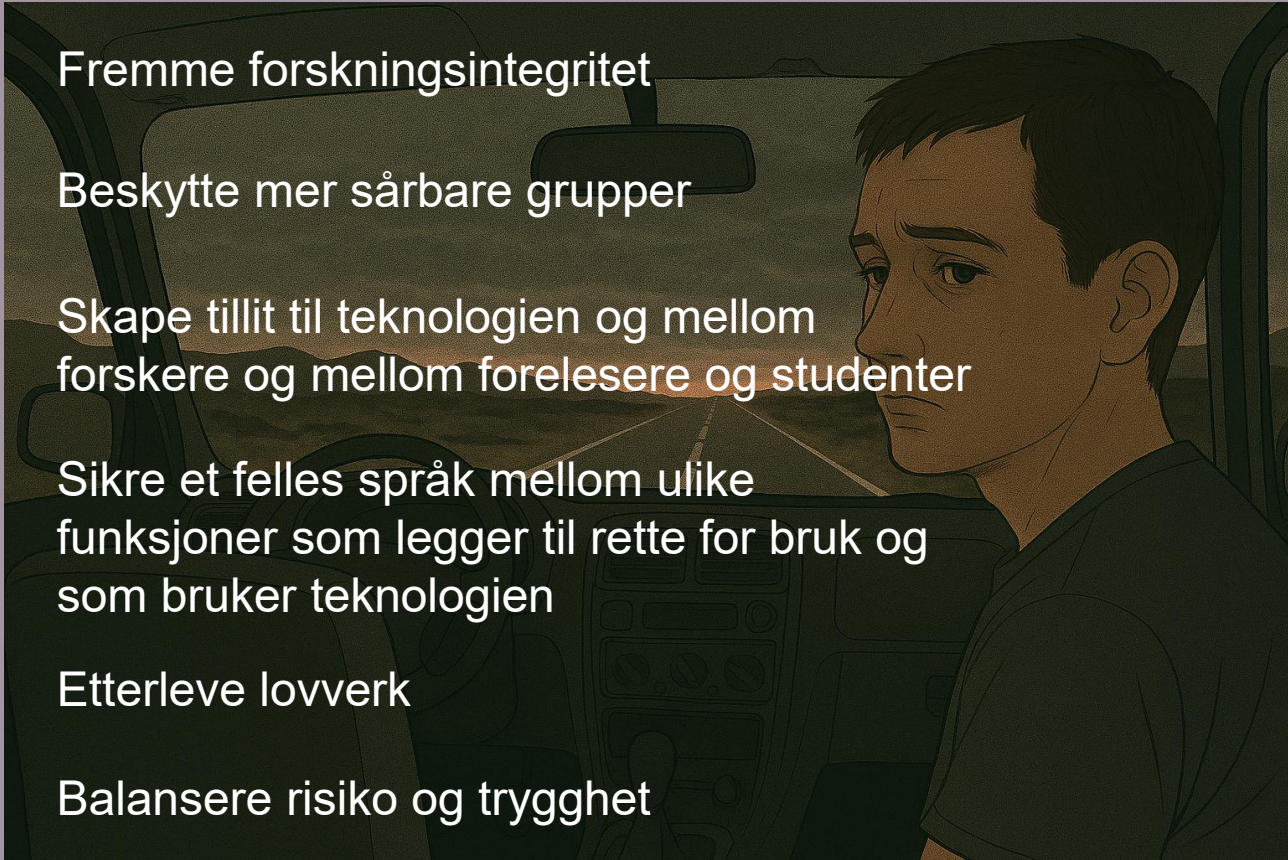
8. Considering implications of AI use is important

10. More than half would look to academic societies for AI guidance

9. Confusion around available guidance



46% said that the institution they work at does not have a policy on AI usage in their work, and an additional 26% said they did not know if there is a policy.



Fremme forskningsintegritet

Beskytte mer sårbare grupper

Skape tillit til teknologien og mellom
forskere og mellom forelesere og studenter

Sikre et felles språk mellom ulike
funksjoner som legger til rette for bruk og
som bruker teknologien

Etterleve lovverk

Balansere risiko og trygghet

Kan felles
retningslinjer
plassere oss i
førersetet?

Ansvarlig KI

=

lovlig, sikker og
etisk forsvarlig*

Lovlig

- Er verktøyene vurdert og godkjent av noen med rett kompetanse?
 - GDPR, personvernlovgivning, KI-forordning
- Bruker jeg verktøyene til det de er godkjent for og i henhold til gjeldende lovverk?
 - KI-fordordning, forvaltningsloven, offentlighetsloven, diskrimineringslovverk mm.

Sikker

- Har vi tilstrekkelig kompetanse til å ivareta sikkerhet?
- Er systemene og algoritmene robuste og trygge i bruk, og har de tilgang til rett data?
- Logges aktivitet og feil?

Etisk forsvarlig

- for å gjøre gode etiske avveininger må vi være klar over hvordan teknologien **fungerer** og hvilke begrensninger den har
- hvilke implikasjoner kan **bruken** få for ulike fagområder og læring hos studenter- er det hjelpsomt?
- overgripende spørsmål knyttet til for eksempel klima og miljø, inkludering i samfunnet, demokrati, maktbalanse

Hva sier EU?

Living guidelines on the **RESPONSIBLE USE OF GENERATIVE AI IN RESEARCH**

ERA Forum
Stakeholders'
document

For forskeren

- Du er ansvarlig for all vitenskapelig output!
- Bruk GenKI transparent
- Vær nøye med personvern, IP og konfidensialitet
- Respekter lovverk
- Sørg for kontinuerlig læring
- Ikke bruk KI til arbeid som kan være sensitivt, som f.eks. peer review

Living guidelines on the **RESPONSIBLE USE OF GENERATIVE AI IN RESEARCH**

ERA Forum
Stakeholders'
document

For forskningsorganisasjonen

- Fremsnakk, veiled og støtt bruken av ansvarlig KI
- Følg aktivt utviklingen i bruk av KI ved institusjonen
- Integrer retningslinjer for KI i generelle retningslinjer for forskning
- Når mulig og nødvendig, implementer lokalt drevne KI-verktøy som ivaretar datasikkerhet og konfidensialitet

Mange gode eksempler på retningslinjer og veiledere/prinsipper fra sektoren



Retningslinjer for bruk av
generativ kunstig intelligens

Type dokument	Retningslinje
Forvaltes av	Seksjon for Digital
Godkjent av	Direktør for Organ
Gjelder fra	03.06.2025
Neste revisjon innen	30.09.2025
Klassifisering	Åpen
Referanse ISO	Ikke vurdert
Referanse NSMs grunnprinsipper for IKT-sikkerhet	
Referanse LOV/Regel	
Referanse interne dokumenter	Retningslinje for b underlagt NTNUs IKT-reglementet



Find Tools & Services

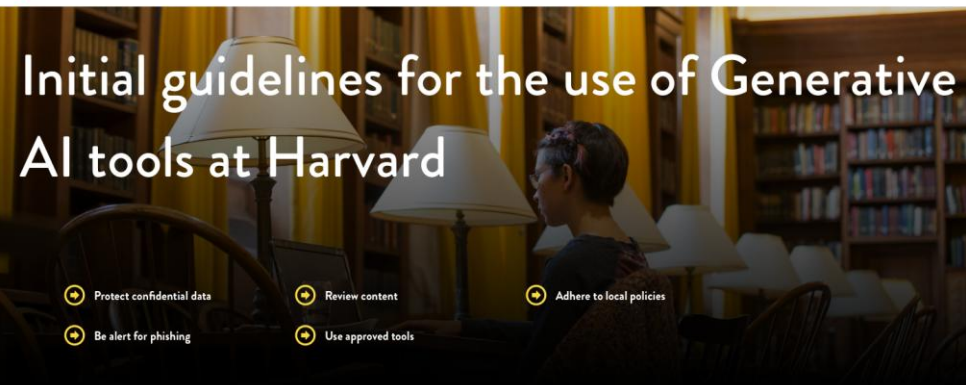
Report a Problem

About IT

Service Status

Get Help

Q



1. Formål

Retningslinjene skal sikre at bruk av IKT-verktøy med gen
lovlig, ansvarlig, etisk riktig og sikker. De baserer seg på r
nasjonale og europeiske lovverk, retningslinjer og anbefalinger.

Formålet med retningslinjen er å sikre at alle som bruker KI-verktøy

- bruker verktøy NTNU har godkjent

Home / Generative Artificial Intelligence (AI) / Generative AI Guidelines

Generative AI is a type of artificial intelligence that can learn from and mimic large amounts of data

Forberede seg

Eksamenslokale



Slik verifiserer du KI-generert innhold (dette ser sensor etter)

ferdigheter. Bruk KI ansvarlig for å

om det er ditt eget, anses som fusk.
alt KI-redigert innhold.

er i hvilken grad du kan benytte KI-
og foreligge skriftlig, f.eks. i Canvas.
erer også eventuelt av KI-bruk på

av eller eksamen.
øy som er godkjent, hvordan du
i-generert innhold.

**«De etablerte
prinsippene er
fremdeles de mest
vesentlige»***

Hallvard Fossheim, UiB

**Pålitelighet
Ærlighet
Respekt
Ansvarlighet**

KI i høyere utdanning



Malthe-Sørenssen-utvalget

Utvalget om kunstig intelligens i høyere utdanning

[Om utvalget](#) [Mandat](#) [Medlemmer](#) [Innspill](#) [Møter i utvalget](#)

Medlemmer



Utvalget skal:

- gi en overordnet analyse av hvordan KI påvirker høyere utdanning
- gi en analyse av hvilke muligheter og utfordringer KI innebærer for høyere utdanning, herunder utforming av studietilbud, undervisning og læringsaktiviteter, eksamen, vurdering, kvalitetsarbeid, studentenes arbeidsvaner og tilegnelse av norsk fagspråk.
- gi råd om hvordan institusjonene og fagmiljøene best kan forholde seg til KI, for å bidra til at fastsatte målsettinger for høyere utdanning blir oppnådd med effektiv bruk av ressurser
- gi råd om hvordan institusjonene og fagmiljøene kan veilede studenter i god KI-bruk
- gi råd til statlige myndigheter om hvordan de bør utnytte mulighetene og løse utfordringer med KI i høyere utdanning og hvilke tiltak som eventuelt bør iverksettes. Råd om endring av regelverk eller statlig styring skal begrunnes med referanse til relevante forarbeider og grunnlagsdokumenter.

Retningslinjer ...hva så?

KI-rådets **mandat**

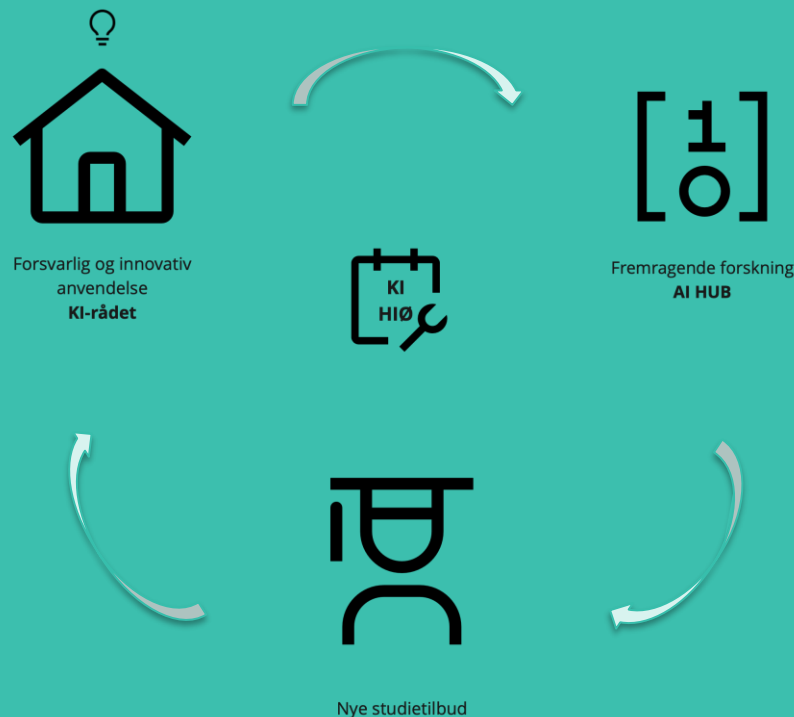
Forankret i rektors ledergruppe

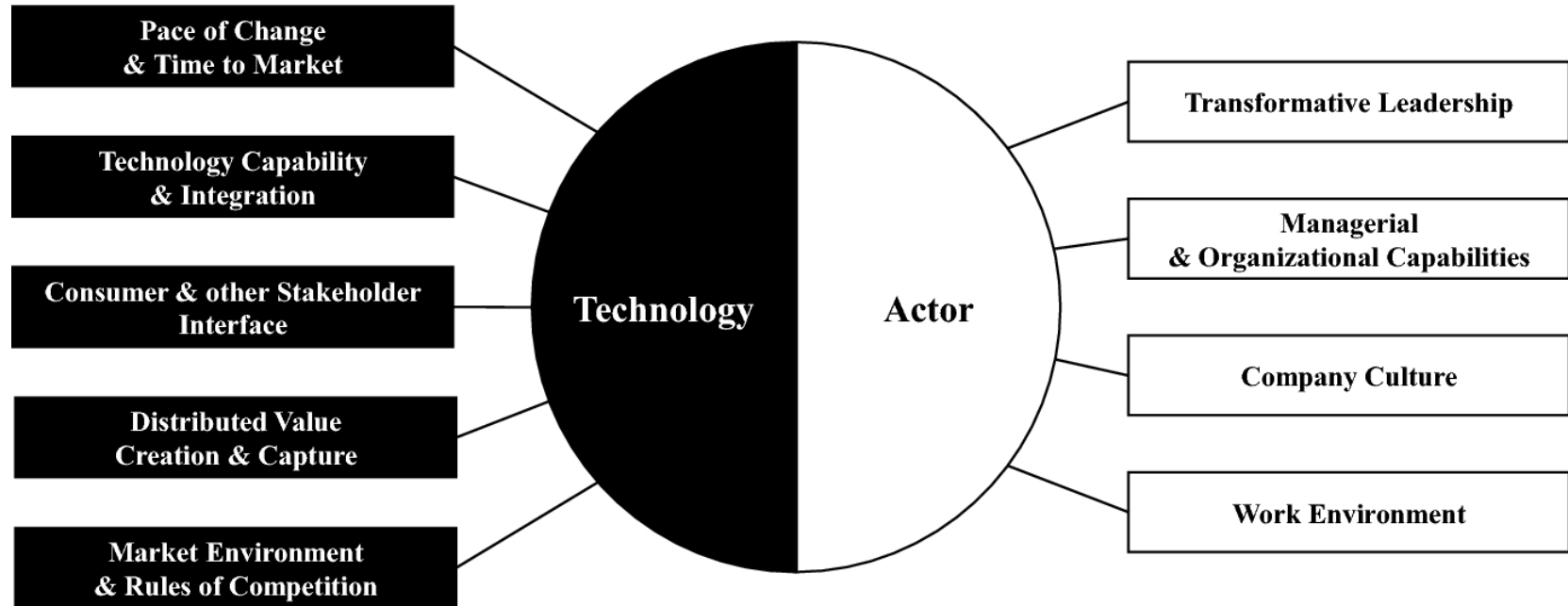
Overordnet målsetning

- › KI-rådet skal legge til rette for *forsvarlig og innovativ bruk* av KI ved Høgskolen i Østfold
- › KI-rådet skal legge til rette for at KI brukes til det beste for samfunnet

Funksjon

- › **Koordinerende**
- › **Rådgivende**
- › **Fasiliterende**





Ansvar og
forankring

Kompetanse-
og
erfaringsdeling

Monitorere
utvikling

Respektere
ulikhet

Tverrfaglig
involvering
tidlig

Tjenester
og
flaksehalser

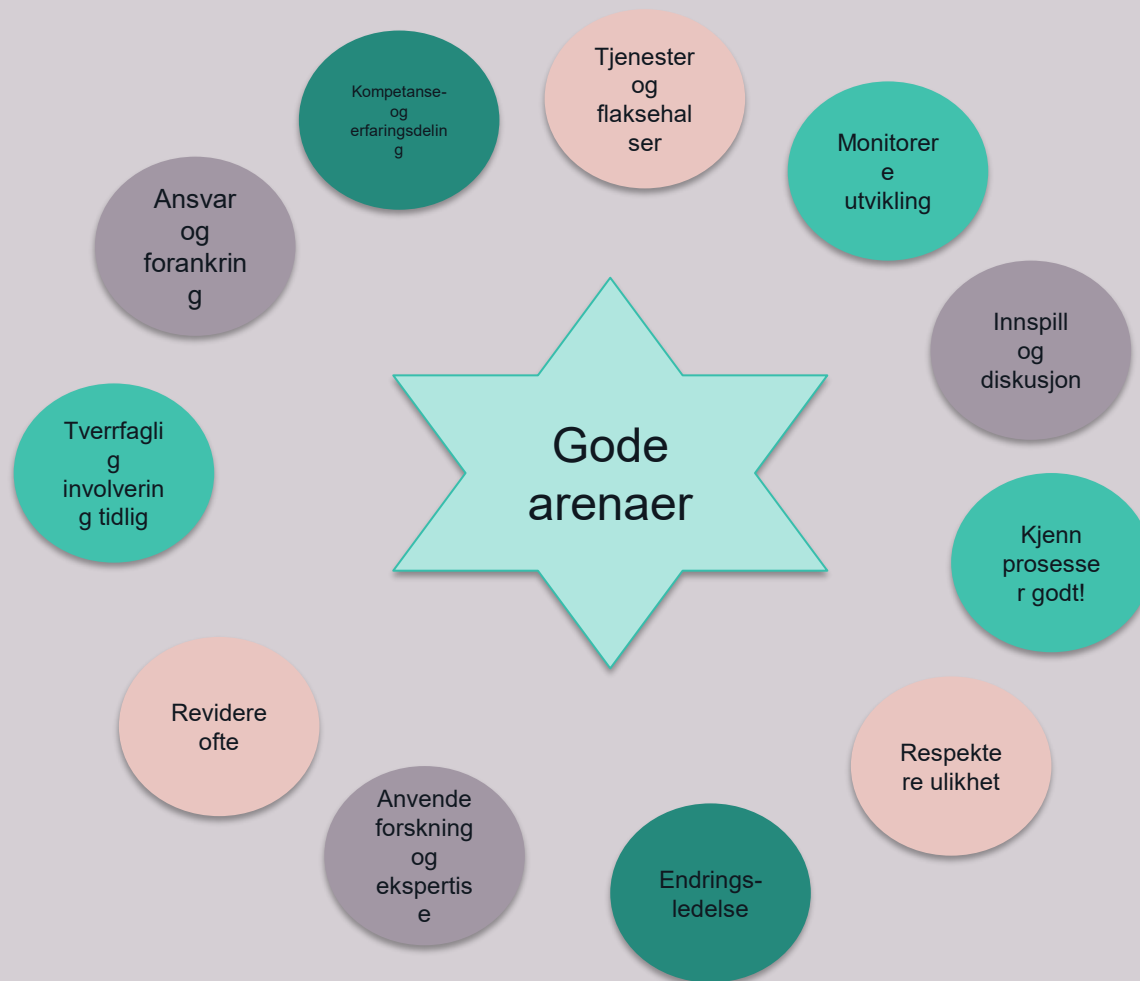
Innspill og
diskusjon

Revidere
ofte

Anvende
forskning
og
ekspertise

Kjenn
prosesser
godt!

Endrings-
ledelse



På hvilken måte endrer KI hvordan vi ser på verden, og hvordan forskyves makten?

Hvordan dannes kunnskap i fremtiden?

Hvilken kunnskap og ferdigheter bør man egentlig ha for å utøve ulike profesjoner, inkludert forskning?

Ja, det er behov for retningslinjer/veiledere som kompass og utgangspunkt for et felles språk, men **samtidig** kontinuerlige refleksjoner rundt hvordan vi ønsker at profesjonene skal utvikle seg!

Vi samarbeider gjerne 😊
Takk for meg!



Referanser

- (1) <https://ai-2027.com>
- (2) https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en?filename=ec_rtd_ai-guidelines.pdf
- (3) <https://www.youtube.com/watch?v=wmQVdzBRnN4>
- (4) <https://www.regjeringen.no/no/dokumenter/fremtidens-digitale-norge/id3054645/?ch=2>
- (5) <https://www.regjeringen.no/no/aktuelt/okt-bruk-av-ki-i-staten/id3097740/>
- (6) <https://economictimes.indiatimes.com/magazines/panache/ais-research-blunder-how-a-mistake-sparked-a-chain-of-flawed-scientific-papers-can-artificial-intelligence-be-trusted-in-academia/articleshow/118821229.cms?from=mdr>
- (7) https://www.reddit.com/r/changemyview/comments/1k8b2hj/meta_unauthorized_experiment_on_cm_v_involving/
- (8) https://www.khrono.no/eksperter-frykter-at-ki-gjor-studentene-dummere/991970?utm_campaign=2025-08-25%3A+Tror+KI+gjor+studentene+dummere+%2F%2F+To+universiteter+på+pallen+%2F%2F+Ikke+forberedt+på+studier&utm_source=newsletter&utm_medium=email&utm_content=header&mm-subscriber=1
- (9) <https://www.khrono.no/i-den-ne-artikkelen-ligger-en-hemmelig-beskjed-slik-skulle-fagfellene-lures/987702>
- (10) <https://www.khrono.no/chatgpt-formidlet-falsk-forskning/994596>
- (11) Studiebarometeret 2024 (x): <https://www.nokut.no/nyheter/fire-av-fem-studenter-bruker-kunstig-intelligens-i-studiene/>
- (12) <https://www.nature.com/articles/d41586-023-02980-0>
- (13) <https://corp.oup.com/news/how-are-researchers-responding-to-ai/>
- (14) <https://www.hiof.no/tjenester/it/ai/retningslinjer/>
- (15) https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en?filename=ec_rtd_ai-guidelines.pdf
- (16) NTNU: <https://i.ntnu.no/documents/1305837853/1306916738/2025-05-06+Retningslinjer+for+bruk+av+IKT-verktøy+med+generativ+KI.pdf/66de9b81-8631-64fe-8ddc-2deca559b043?t=1749022115156>
- (17) https://uit.no/eksamen#modal_811784
- (18) <https://www.huit.harvard.edu/ai/guidelines>
- (19) <https://www.forskningsetikk.no/ressurser/fbib/forskning-samfunn/kunstig-intelligens/>
- (20) <https://malthesorensenutvalget.no>
- (21) Nadkarni, S. & Prügl, R (2021). Digital transformation: a review, synthesis and opportunities for future research.