Norges miljø- og biovitenskapelige universitet

Institutt for datavitenskap

Digitalisering på Ås

European standardization efforts from FAIR toward explainable-AI-ready (XAIR) data documentation in materials modelling

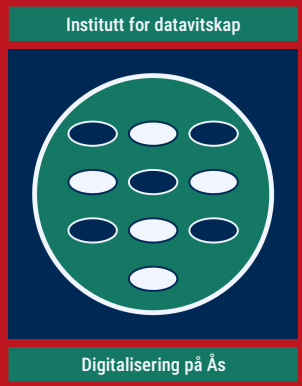
Martin T. Horsch,[1] Björn Schembera,[2] Heinz A. Preisig[3]

[1]Norwegian University of Life Sciences, Department of Data Science
[2]University of Stuttgart, Institute of Applied Analysis and Numerical Simulation
[3]Norwegian University of Science and Technology, Department of Chemical Engineering

Fakultet for realfag og teknologi

Institutt for datavitenskap

Noregs miljø- og biovitskaplege universitet



1. The need
2. The state of the art
3. What are we proposing?

Epistemic opacity (Humphreys, 2011): A cognitive "process is **epistemically opaque** relative to a cognitive agent *X* at time *t* just in case *X* does not know at *t* all of the **epistemically relevant elements** of the process."

European AI Act proposal: "To address the **opacity** that may make certain AI systems **incomprehensible to or too complex for natural persons**, a certain degree of transparency should be required for high-risk AI systems. […] High-risk AI systems should therefore be accompanied by **relevant documentation**".

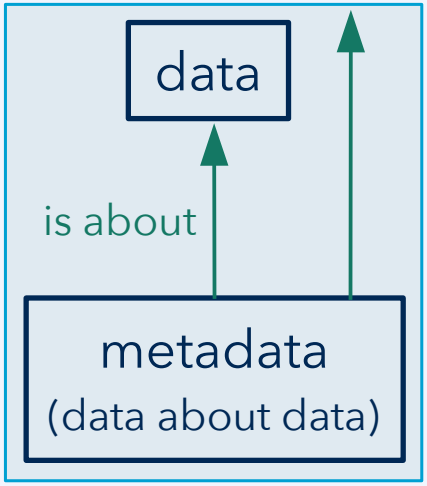
Digitization and digitalization

Metadata are "descriptive data about an object" (ISO 11179).

Digitization

Digitalization

Industry 5.0



FAIR digital object

Leiden 2022 Declaration for **FAIR digital objects**:

https://www.fdo2022.org/site/fdo/programme/leiden-declaration

The librarian:

- Focus on **archival** and **curation**
- **Help humans** to make FAIR use of digital artefacts
- Focus on **provenance, like for artefacts in a museum**, so humans understand where they come from

The engineer:

- **Computers must "understand"** what the digital artefacts mean
- **Focus on knowledge**/meaning
- We need **FAIR digital objects**

Norwegian University of Life Sciences

Epistemic metadata

Metadata are "descriptive data about an object" (ISO 11179).

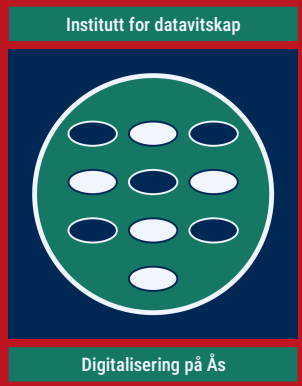
Epistemic metadata are those that help establish the knowledge status of data.

> **Epistemic metadata:**
>
> **a)** "what **knowledge claim φ** has been formulated?,"
> **b)** "where do the data and the claim come from?" (**provenance**),
> **c)** "what **validity claim** was made about φ?,"
> **d)** "why should we accept any of this?" (**grounding**).

Case study from molecular thermodynamics

- First stage, evaluating ten journal articles, doi:10.5281/zenodo.7516532.
- Second stage, discussing twelve claims, doi:10.5281/zenodo.7608074.

Noregs miljø- og biovitskaplege universitet



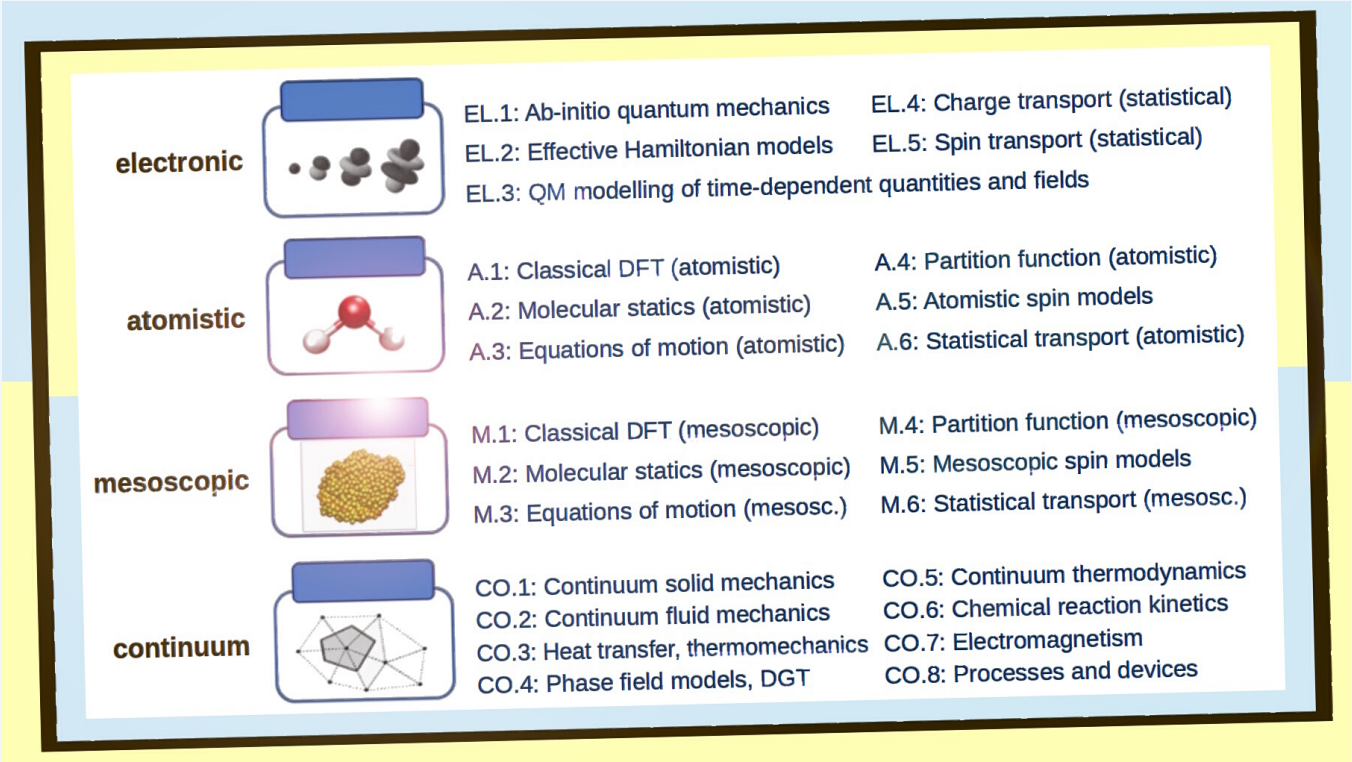
1. The need
2. The state of the art
3. What are we proposing?

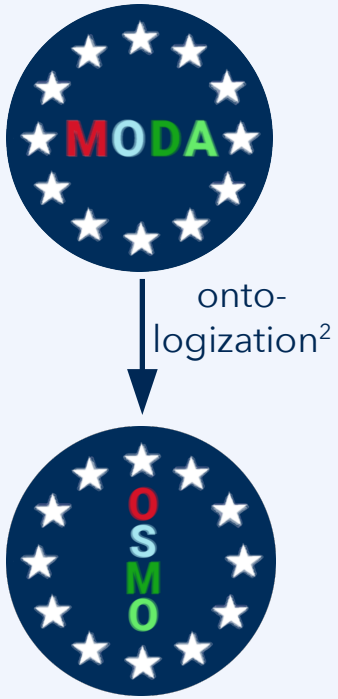
EMMC-related development efforts from 2017 onward (stable release soon) have lead to a novel, radically physicalistic **top-level ontology**: The **Elementary Multiperspective Material Ontology** (EMMO). The EMMO includes a **Peircean semiotics** as a "perspective" on cognition.

The **European Materials Modelling Council** (EMMC) community has developed three **CEN workshop agreements** (CWAs) as documentation standards: CEN 17284 **MODA** ("model data" **provenance**), CWA 17815 **CHADA** ("characterization data" **provenance**), CEN 17960 **ModGra** ("model graphs" for **process model toplogies**).

RoMM (2017), MODA (2018), and CHADA (2021)

As an attempt at metadata standardization,[1, 2] RoMM/MODA resulted in a closed epistemic space with a rigid categorization of modelling methodologies. MODA/CHADA documentations are hard to create and **hard to use by humans and not machine-actionable**.





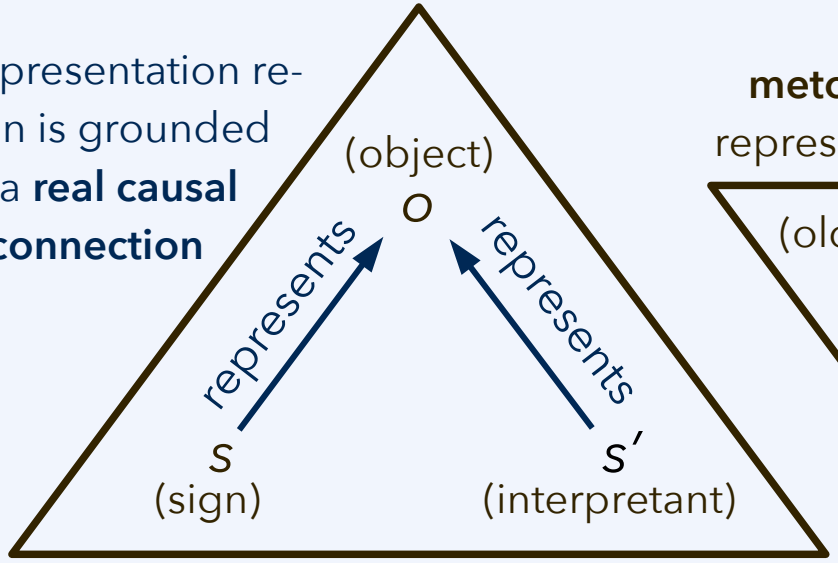
[1]A. F. de Baas (ed.), *What Makes a Material Function?*, ISBN 978-92-79-63185-6, **2017**.

[2]*Journal of Chemical & Engineering Data* 65, 1313, doi:10.1021/acs.jced.9b00739, **2020**.

EMMO[1] and Peircean semiotics

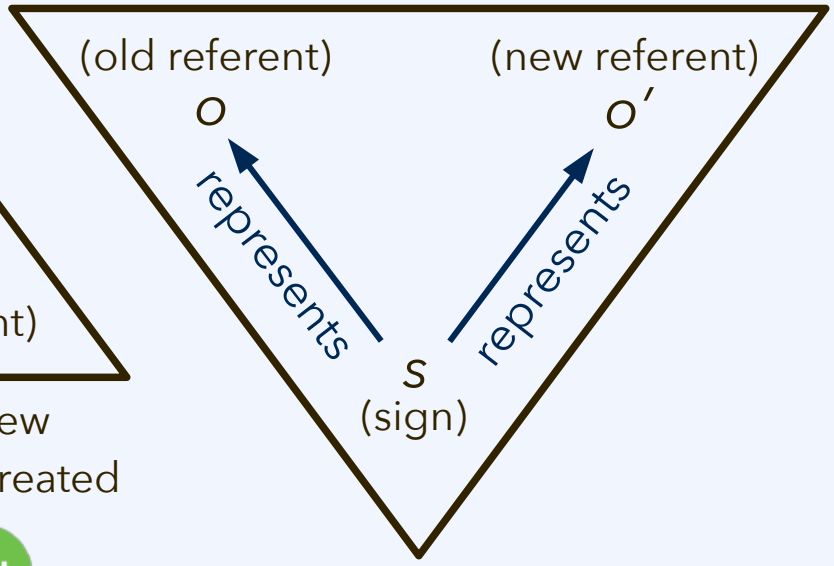
Peircean semiotics: By using a sign (1st) for an object (2nd), a "Third" is created.

the representation relation is grounded in a **real causal connection**



semiosis, a process by which a new representamen, the interpretant, is created

metonymization, a process by which a representamen is assigned a new referent



C. S. Peirce

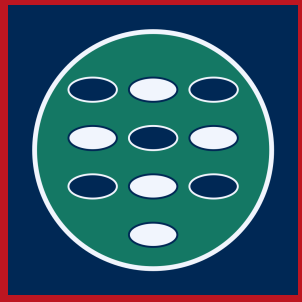
Elementary Multi-perspective Material Ontology (EMMO)

The EMMO[1] combines this with mereotopology – foundational ontology as **mereosemiotics**.

[1]The work on the EMMO (2017 – present) is coordinated by Emanuele Ghedini.

Noregs miljø- og biovitskaplege universitet

Institutt for datavitskap

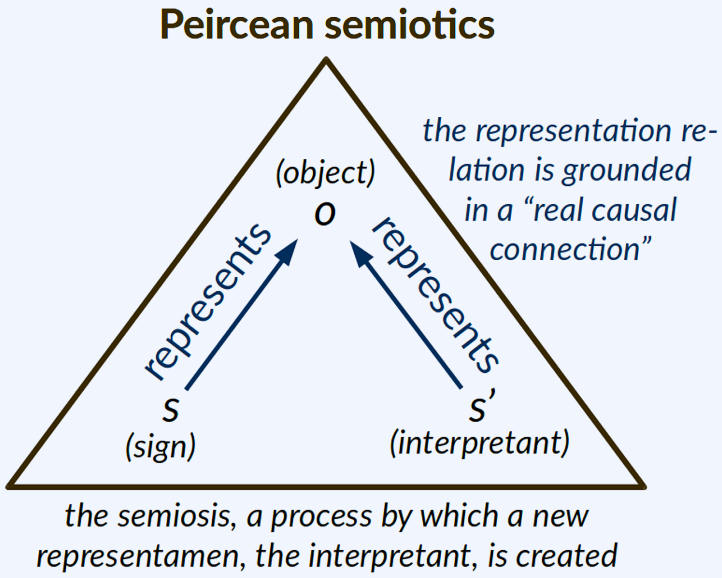


Digitalisering på Ås

1. The need
2. The state of the art
3. What are we proposing?

The **PIMS-II mid-level ontology** implements a data documentation strategy based on **epistemic metadata** building on Peircean semiotics. Our present work has its focus on **knowledge claims** (what we know from data) and their assessment through validity claims, including **reproducibility claims**.

Peircean semiotics: Provenance



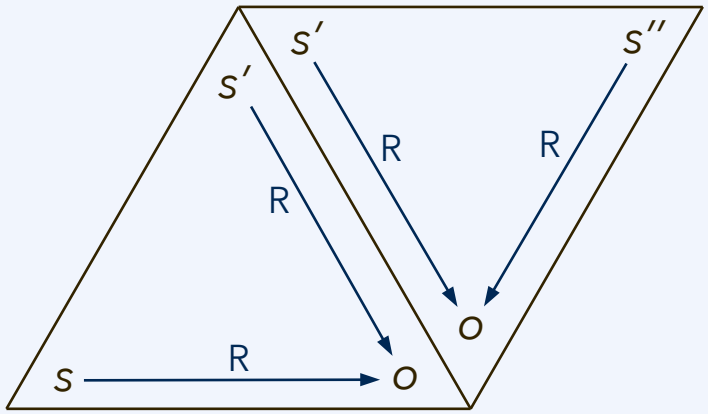
Cognitive process (example):

- First, experimental data *s* for material *o* are used to parameterize a model, obtaining model *s′*.
- Then, a simulation is done using model *s′*, yielding the simulation result *s″* (which also represents *o*).

Research workflows as cognitive processes:

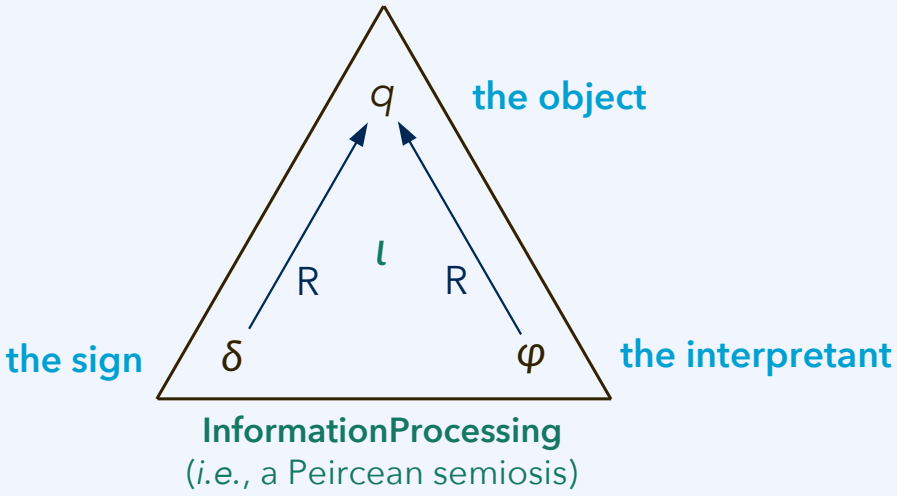
Each cognitive step starts from one representation relation, *e.g.*, *Rso*, and creates a new one, *Rs′o*.

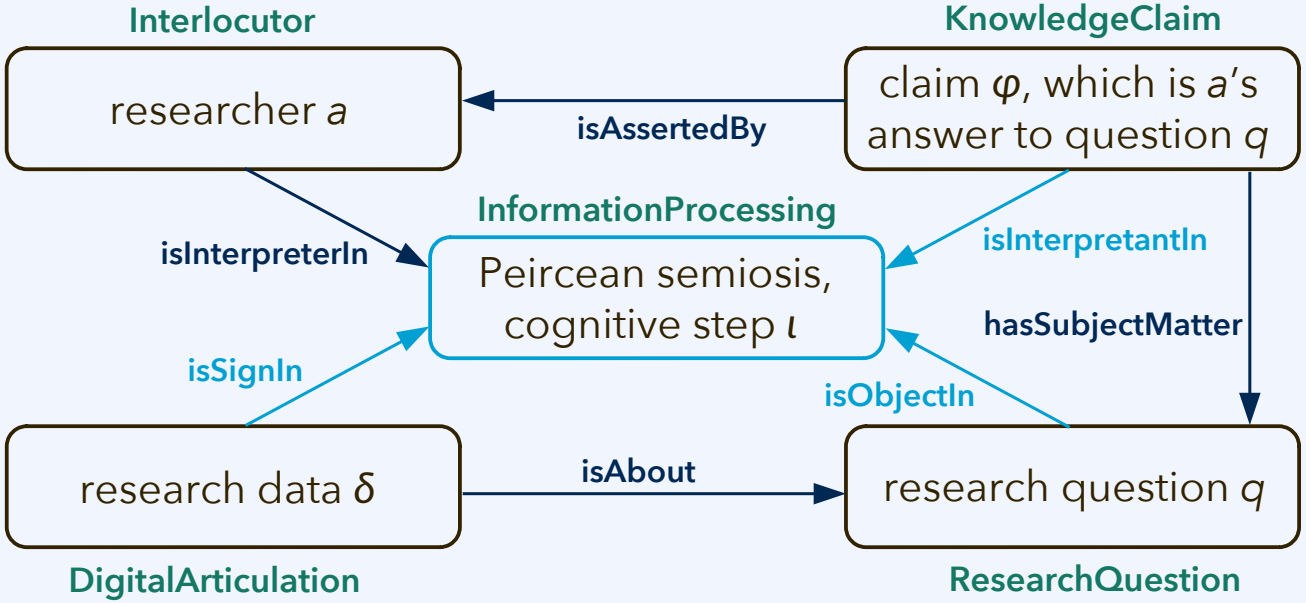
The successor step reuses *Rs′o* and creates the next relation, *Rs″o*.



cognitive process *κ*

Peircean semiotics: Knowledge claims





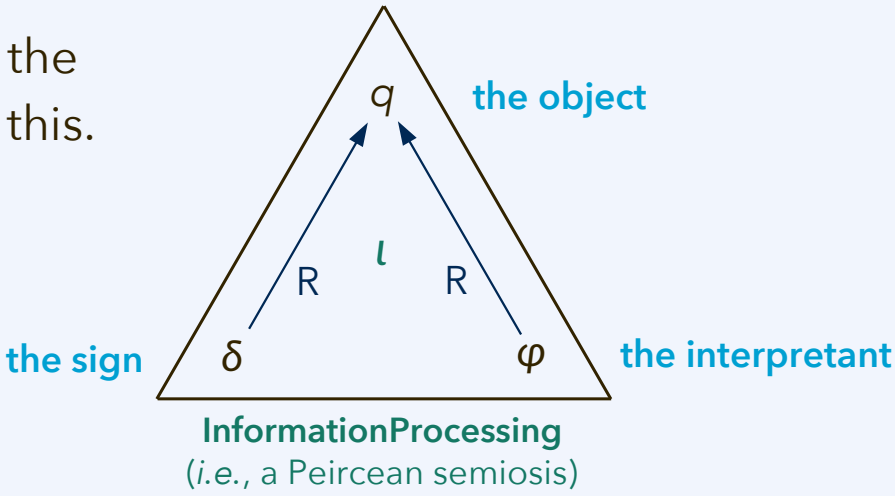
Peircean semiotics: Knowledge claims

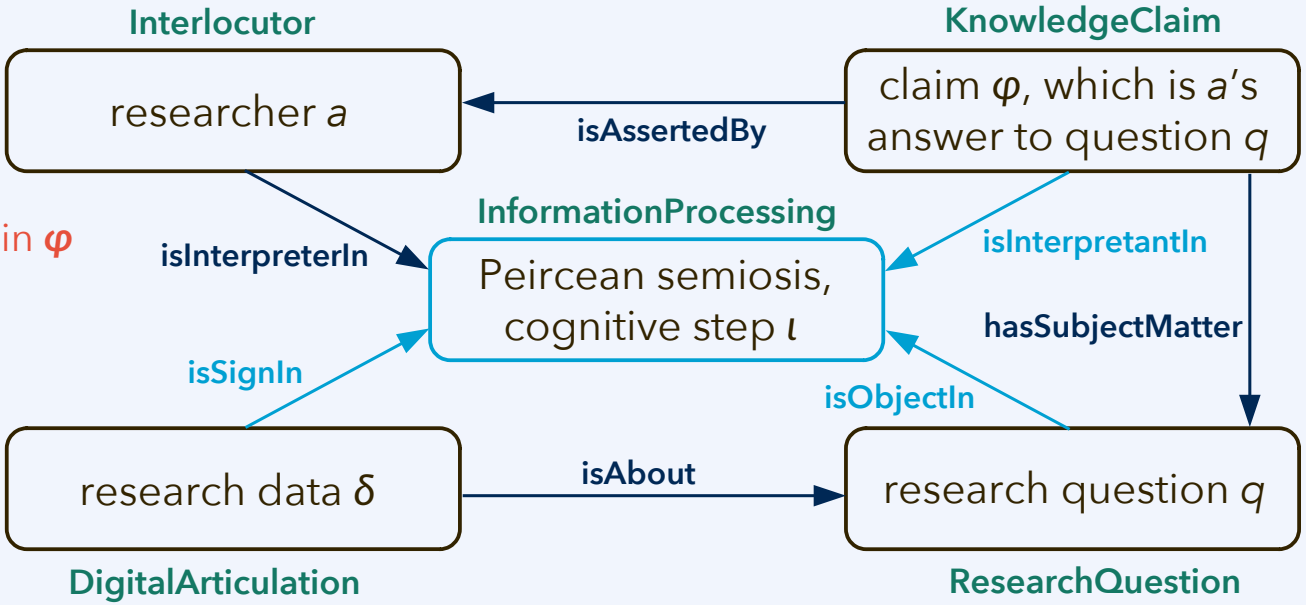
Researcher: I did *κ*, obtained the data *δ*, and found out *φ* from this. **Now I claim to know *φ*.**

Data infrastructure: Therefore, we know that **it is possible to do *κ* and find *φ*.**

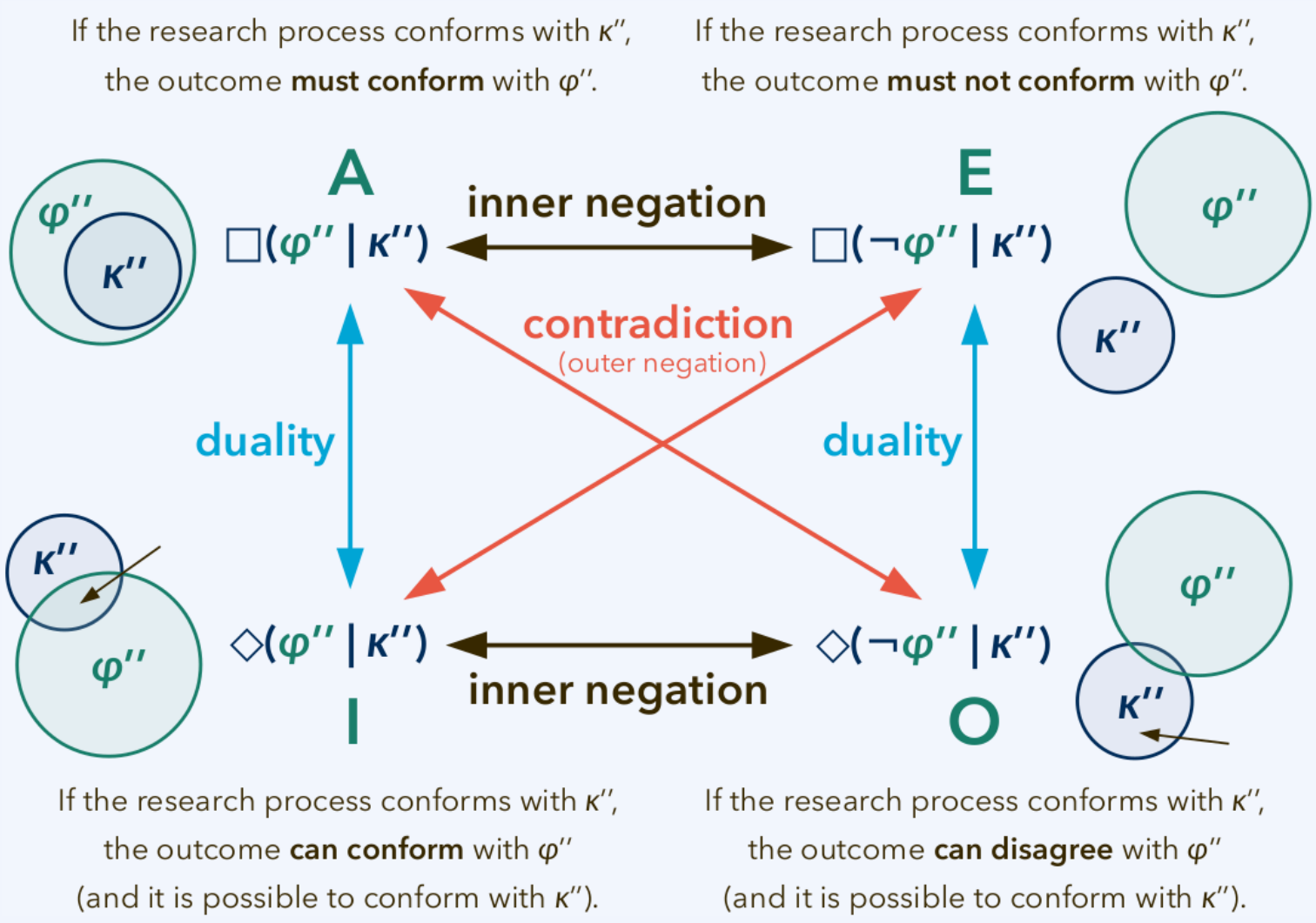
$\Diamond(\varphi \mid \kappa)$

Read: "**possible** to obtain *φ* **given** that *κ* was done."





Modal square of opposition



Reproducibility claims

Common formulation and schema for reproducibility claims (RCs):

«Whenever research process κ'' is carried out, it must lead to the outcome φ''.»

1) Reseacher *a* did κ and found φ.

   Here, *a* also made the **positive reproducibility claim $\psi = \Box(\varphi'' \mid \kappa'')$**.

2) Reseacher *b* did γ, **consistent with κ''**, and found ζ, **inconsistent with φ''**.

   Here, *b* made the **negative reproducibility claim $\Diamond(\neg\varphi'' \mid \kappa'') \equiv \neg\Box(\varphi'' \mid \kappa'') \equiv \neg\psi$**.

3) What is relevant there is the **contradiction between ψ and $\neg\psi$**.

provenance metadata κ provenance paradata κ'	knowledge claim metadata φ knowledge claim paradata φ'
provenance orthodata $\kappa'' = \kappa - \kappa'$	**knowledge claim orthodata $\varphi'' = \varphi - \varphi'$**
«repeat κ, but no need to retain κ'»	«obtain φ again, except for φ' maybe»

Norges miljø- og biovitenskapelige universitet

Institutt for datavitenskap

Digitalisering på Ås

European standardization efforts from FAIR toward explainable-AI-ready (XAIR) data documentation in materials modelling

Martin T. Horsch,[1] Björn Schembera,[2] Heinz A. Preisig[3]

[1]Norwegian University of Life Sciences, Department of Data Science
[2]University of Stuttgart, Institute of Applied Analysis and Numerical Simulation
[3]Norwegian University of Science and Technology, Department of Chemical Engineering

Fakultet for realfag og teknologi

Institutt for datavitenskap