Old AI Meets New AI in the Logic of Scientific Discovery

Ioannis Votsis
Associate Professor, Philosophy
Research Head, Reimagining HE in the Age of AI
Northeastern University – London
ioannis.votsis@nulondon.ac.uk
www.votsis.org

AAAI23 Spring Symposium Series, 27th-29th March 2023

Talk plan

(1) Old AI vs. New AI

(2) Hybrid Approaches

(3) A Proposed Hybrid Approach

(4) Adding or Removing Content

(5) Two Constraints on Theory Choice

Old AI vs. New AI

What is old and what is new AI?

- **Old AI**: Roughly, the computational implementation of logical inferences to process symbolic representations.

  Examples: expert systems, automated theorem provers, computational argumentation.

- **New AI**: Roughly, the computational implementation of statistical inferences to process neural representations.

  Examples: shallow and deep neural nets (supervised, unsupervised and reinforcement learning).

Who has the (relative) upper hand?

Positive characteristics	Old AI	New AI
Adaptiveness		☝
Compositionality	☝	
Data efficiency	☝	
Detecting patterns		☝
Formal verification	☝	
Interpretability	☝	
Learning from data		☝
Reasoning	☝	
Simpler expressions	☝	
Universality (domain neutral)	☝	
Unstructured data		☝

Hybrid Approaches

Neuro-symbolic systems: Learn and reason

- The popularity of hybrid, a.k.a. 'neuro-symbolic', approaches has been on the rise in recent years:

  Arabshahi et al. (2021); Garcez et al. (2019); Hamilton et al. (2022); Schockaert & Gutiérrez-Basulto (2022).

- "The aim here is to [integrate] the two most fundamental aspects of intelligent cognitive behavior: **the ability to learn from experience**, and **the ability to reason from what has been learned**" (Valiant 2003: 97).

- Analogies have also been drawn with dual process theories in psychology (Kahneman 2011; Rossi 2022).

Characterising neuro-symbolic systems

- How are the two approaches integrated?

  "In neural-symbolic computing, knowledge is represented in symbolic form, whereas learning and reasoning are computed by a neural network" (Garcez et al. 2019: 2).

- As a general characterisation, this seem a little narrow. Kautz (2020) proposes *five* different ways to integrate them:

  1. Neural net that processes symbols-to-vectors-to-symbols.
  2. Symbolic problem-solver with neural pattern subroutine.
  3. Neural net trained on symbolic rules (input-output pairs).
  4. Symbolic reasoner being fed cascades from neural nets.
  5. Embedding symbolic reasoning into neural nets.

A typology

- Several ways to conceptually integrate (not necessarily by preserving) the neural and symbolic approaches.

- They seem to fall under three types:

  (A) Adapting neural systems to perform symbolic tasks like problem-solving and reasoning (K3; K5).

  (B) Adapting symbolic systems to perform neural tasks like feature extraction and pattern recognition.

  (C) Chain neural and symbolic systems together to coordinate their activity (K2; K4).

A Proposed Hybrid Approach

Implication-driven neuro-symbolic approach

- In a nutshell, the approach suggested here seeks to:

  (1) extract symbolic representations (particularly logical formulae) from neural nets and other sources

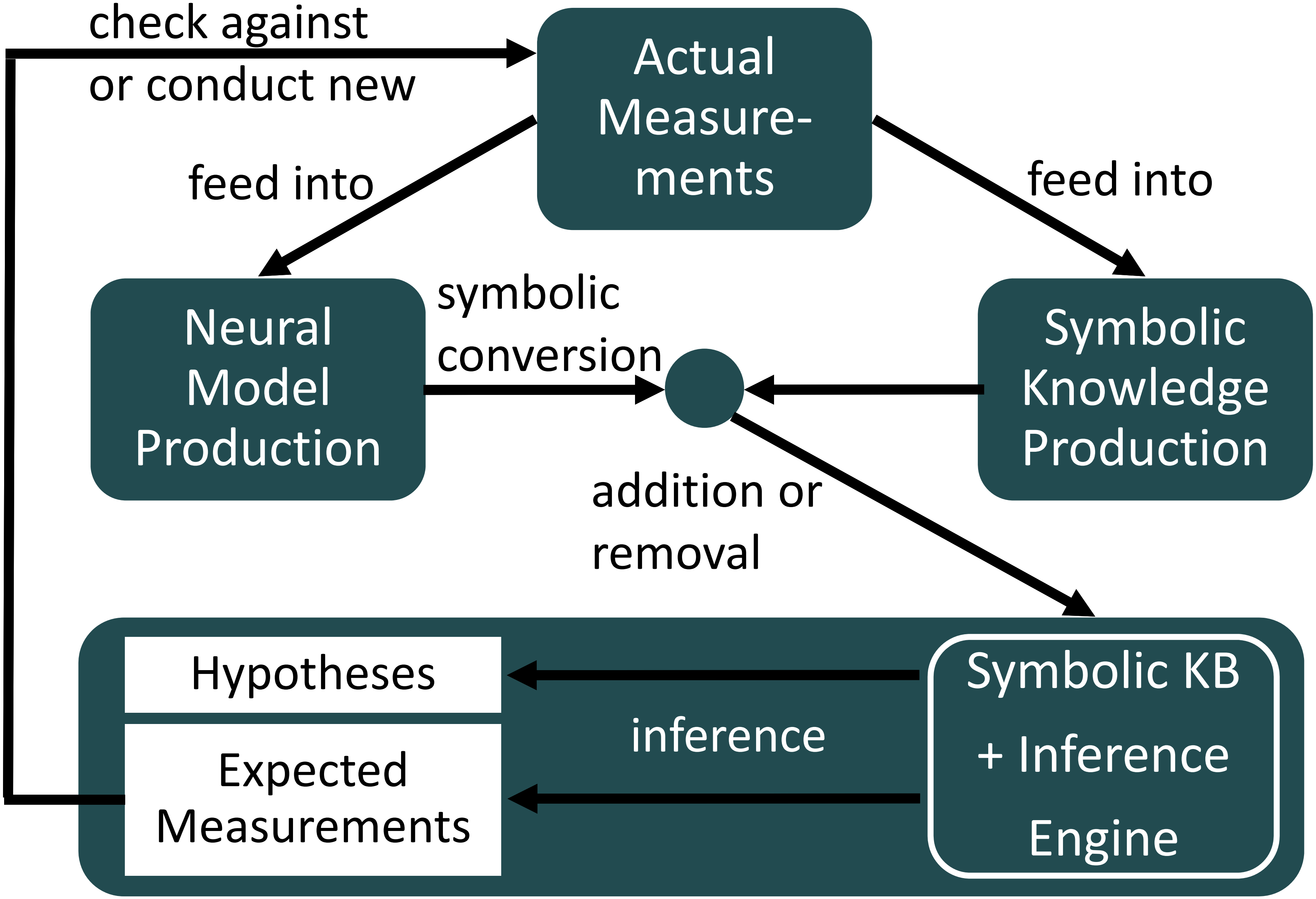
  AND

  (2) process those representations + existing ones using an automated theorem prover

  NB: As such, the approach falls under type C above.

- Besides playing to each tradition's strengths, it allows us to perform all sorts of *implication-driven discovery tasks*.

Diagrammatic form



Extracting symbolic representations

- Some methods that can help with such extraction as well as extraction from other sources (e.g. natural language):

  > **Autoformalisation**: Translating informal (math) proofs into formal proofs (Wu et al. 2022).

  > **Computational argumentation**: Converting neural nets to argument maps (Čyras et al. 2021).

  > **Knowledge Repr. in NNs**: Reversing rule-based and formulae-based translations (Garcez, Gabbay & Broda 2002).

Why reasoning? Why automated theorem proving?

- Arguably, all scientific activity can be *reconstructed* in terms of reasoning <u>and</u> (nearly*) all reasoning can be automated.

- Automated theorem provers (ATP) have been at the forefront of such automation since 50s and have gotten very efficient.

  Applications: logic programming, SAT solvers, formal verification, math proofs.

- Logic Systems: classical (propositional, first-order, higher-order, etc.), non-classical (modal, default, relevance, etc.)

Some implication-driven discovery tasks

Theory modification (removing content to avoid falsities):

<u>From</u>: $T_i \vDash O_j$ where O_j is False. <u>To</u>: $T_i' \nvDash O_j$

Theory modification (adding content to gain truths):

<u>From</u>: $T_i \nvDash O_j$ where O_j is True. <u>To</u>: $T_i' \vDash O_j$

Theory generation (via joint consequence):

<u>From</u>: $T_i \nvDash T_k$; $T_j \nvDash T_k$ <u>To</u>: $T_i \wedge T_j \vDash T_k$

Expected measurement generation (via joint consequence):

<u>From</u>: $T_i \nvDash E_k$; $T_j \nvDash E_k$ <u>To</u>: $T_i \wedge T_j \vDash E_k$

The black box conundrum: Et tu, ATP?

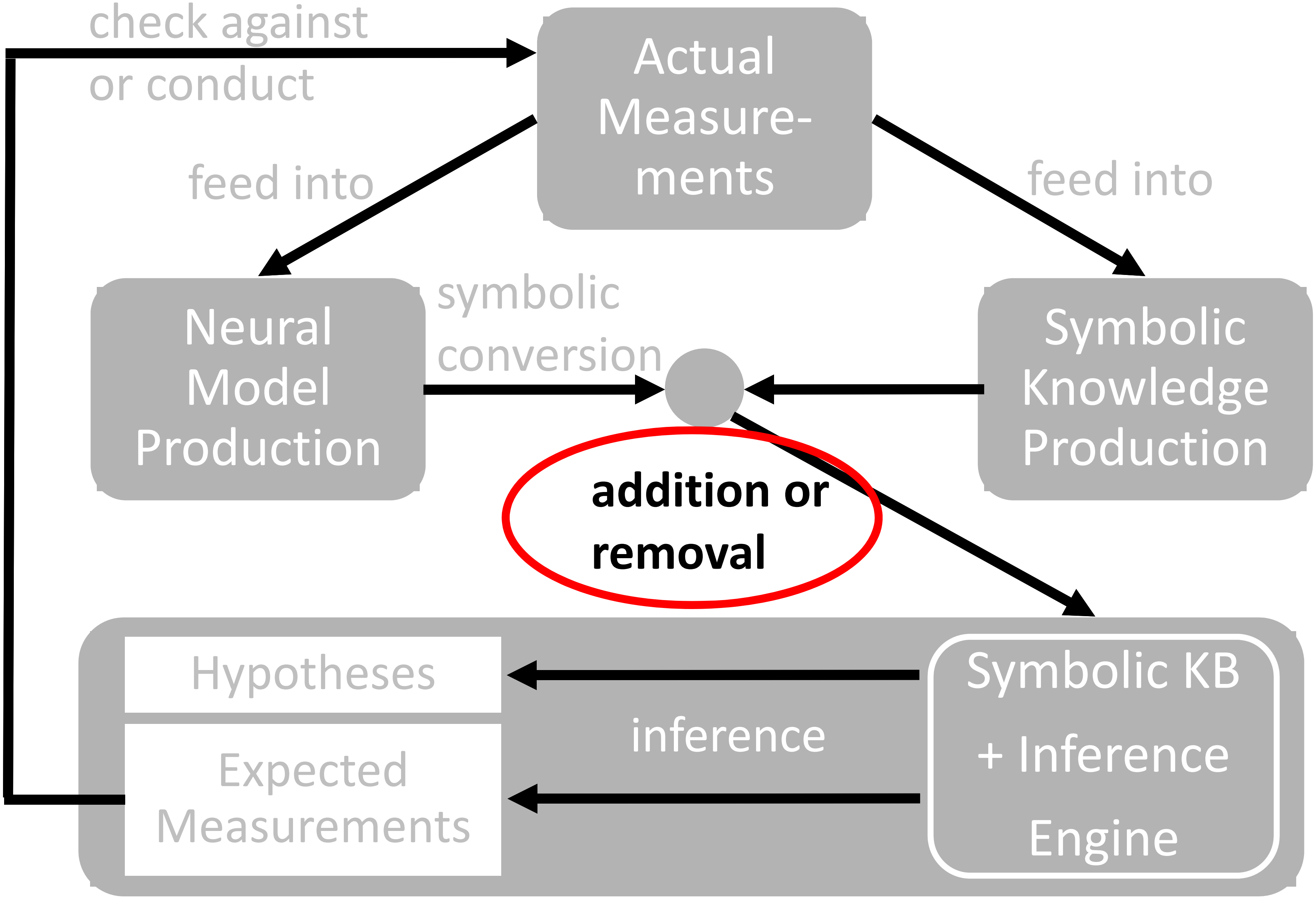
- If we are to use such tools as assistants in scientific discovery, we need human-readable output.
- The trouble with the most widely used ATP method, viz. resolution, is that it sacrifices human readability for efficiency.
- A more suitable tool would be to use a natural deduction (ND) ATP (Pelletier 1998).

  NB: I'm currently trying to develop a hybrid resolution-ND method that translates more easily into ND proofs.

- That ND is more intuitive (at least as a starting point) is also experimentally suggested in Votsis & Nagle (under review).

Adding or Removing Content

Diagrammatic form



Content weakening and content strengthening

- Any theory change (including from no theory to some theory) can be modelled as an addition or deletion of content.

- Two quasi-logical notions (Votsis forthcoming) can help here:

  A theory T is ***content-weakened*** to a theory T^- if and only if $Ded_N(T^-) \subset Ded_N(T)$.

  A theory T is ***content-strengthened*** to a theory T^+ if and only if $Ded_N(T) \subset Ded_N(T^+)$.

- Analogous to BRT (Alchourrón, Gärdenfors & Makinson 1985; Rose & Langley 1986) but w/a restricted consequence notion.

Example: Fresnel to Maxwell

- Fresnel's wave theory of light posits a luminiferous ether to explain phenomena (e.g. reflection and transmission of light).
- We can content-weaken Fresnel's theory by removing the ether assumption and any residual sentences depending on it.
- We can also content-strengthen the theory to an ether-less electromagnetic field.
- That means adding content that construes light:

  * as a vibration in the electric and magnetic field strengths
  * as one of many forms of electro-magnetic radiation

Two Constraints on Theory Choice

Constraints on content

- We have not addressed the crucial question of how to decide which content to add or delete.
- Needless to say, we need to turn to heuristics to make headway on this problem.
- Besides the usual heuristic constraints, e.g. opting for simpler models, we propose two others:

  (1) structural correspondence

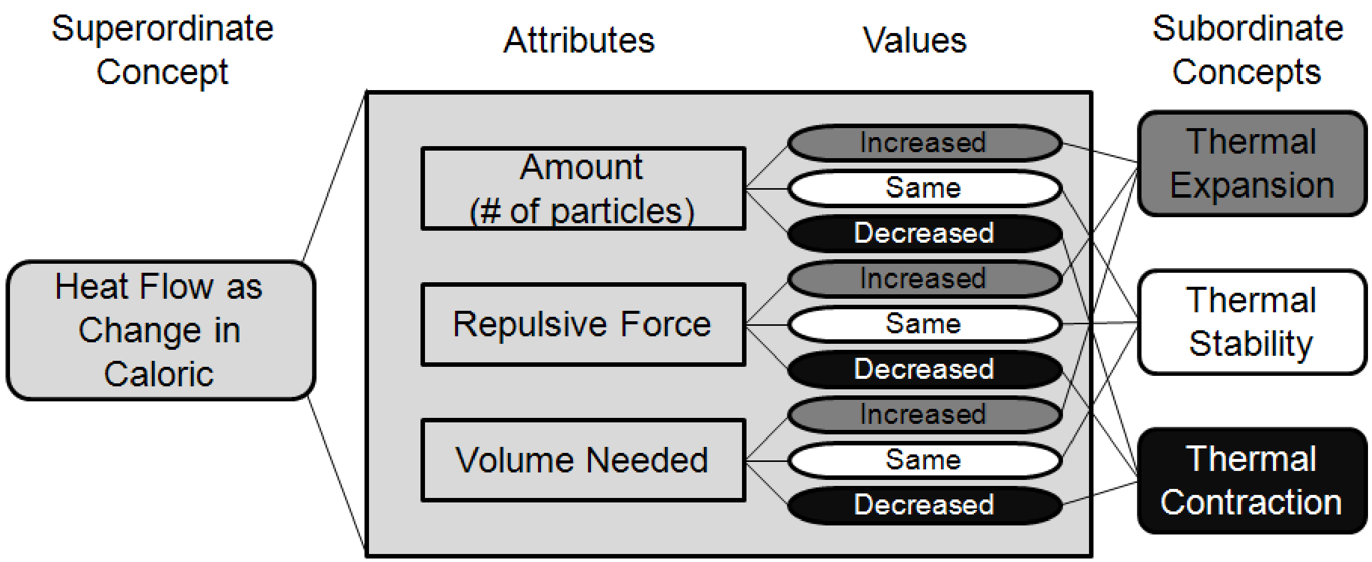
  (2) multiple testing ground consilience

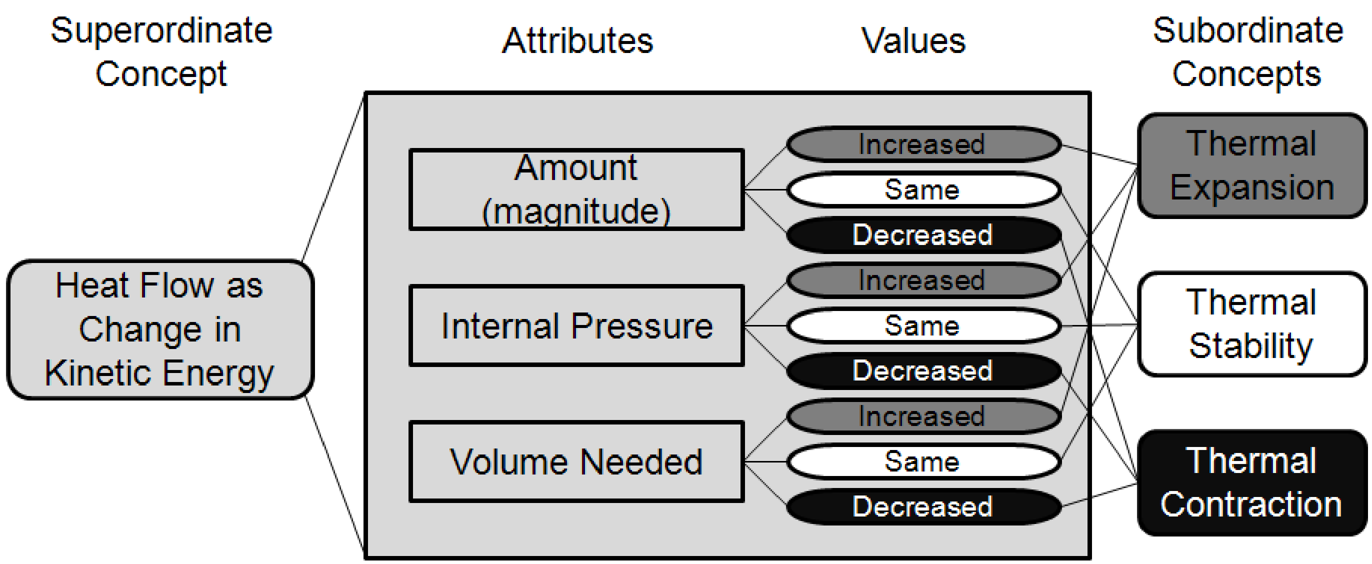
1. Structural correspondence

- Such constraints flow from a view known as ‘structural realism’ (Poincaré 1905; Russell 1927).
- **Structural realism**: Scientific theories (in natural science) describe the unobservable world only up to isomorphism.
- **Structural correspondence**: Any new theory must structurally correspond (at least in some limit form) to the well-confirmed parts of its predecessor.

  * wave theory of light $\hookrightarrow_*$ electromagnetic theory (Worrall 1989)
  * phlogiston theory $\hookrightarrow_*$ oxygen theory (Schurz & Votsis 2014)
  * caloric theory of heat $\hookrightarrow_*$ kinetic theory (Votsis & Schurz 2012)

Caloric and kinetic theories of heat





2. Multiple testing ground consilience

- How should we attribute blame/credit to theories in light of disagreement/agreement with empirical results?

- Suppose:

Central theory:	T_1
Auxiliary hypotheses:	A_1, A_2, A_3
System:	$S_1 \leftrightarrow (T_1 \wedge A_1 \wedge A_2 \wedge A_3)$
Predicted measurements:	$S_1 \vDash O_1 \wedge O_2$
Actual measurements:	O_1 is False; O_2 is True

- **Puzzle**: Given O_1 is False, which part(s) of S_1 are needed and which must be replaced or at least removed?

Example

- **Step 1**: Check if all parts of S_1 are needed to derive O_1.

$T_1 \wedge A_1 \wedge A_2 \vDash O_1$ Defeasibly learn: A_3 ✓

- **Step 2**: Check if remaining parts fare well in other testing grounds.

$T_1 \wedge A_4 \vDash O_3$; O_3 is True Defeasibly learn: T_1 ✓
$T_2 \wedge A_1 \vDash O_4$; O_4 is True Defeasibly learn: A_1 ✓
$T_3 \wedge A_2 \vDash O_5$; O_5 is False Defeasibly learn: A_2 ✗

- **Step 3**: Weaken A_2 to check if some content can be salvaged.

$T_3 \wedge A_2' \nvDash O_5$; $T_1 \wedge A_1 \wedge A_2' \nvDash O_1$
$T_1 \wedge A_1 \wedge A_2' \wedge A_3 \vDash O_2$ Defeasibly learn: A_2' ✓

- **Step 4**: Strengthen A_2' to check if new content is beneficial.

As above but also $T_4 \wedge A_2'' \vDash O_6$; O_6 is True Defeasibly learn: A_2'' ✓

Summary

- A crude but hopefully useful overview of each tradition's (neural vs. symbolic) strengths and weaknesses was given.
- The subject of hybrid approaches to AI was then broached, and several different variants identified.
- A proposal was made for such a hybrid approach, extracting symbolic repres. from NNs and using ATP to process them.
- Part and parcel of this proposal is the treatment of theory evolution in terms of content addition and/or deletion.
- Two useful heuristic constraints were then discussed: structural corresp. and multiple testing ground consilience.

The End