Accountability of Logic in the New Age of Artificial Intelligence

David Pearce, Universidad Politécnica de Madrid

August 29, 2022

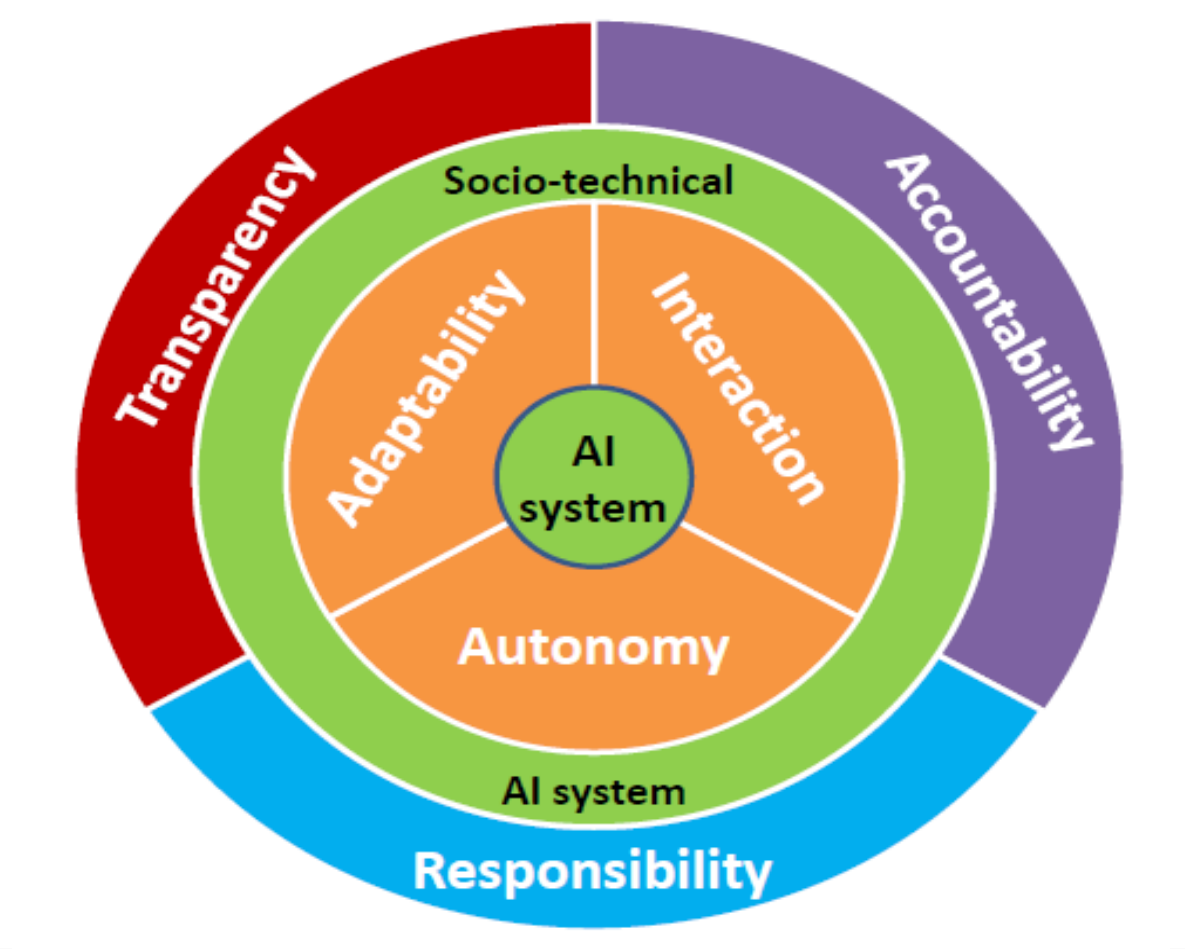
Who cares? New pragmatism in Logic?

We demonstrated that the properties of subjective constraint monotonicity and epistemic splitting can in general be too strong and may exclude some intuitively desired world view for some epistemic programs. We also demonstrated that the foundedness property is not effective in characterizing the well-supportedness of world views from a classical logic perspective and may also exclude some intuitively desired world views. (Shen and Eiter)

And then, after 25 pages of arguments ...

Our approach is based on classical logic, . . . It is upon the users to choose between classical logic based approaches and equilibrium logic based ones for their specific application scenarios.

ART principles for Responsible AI



ART principles for Responsible AI

Accountability: the requirement for the system to be able to explain and justify its decisions to users and other relevant actors.

Responsibility refers to the role of people themselves in their relation to AI systems. The whole socio-technical system in which the system operates, encompassing people, machines and institutions

Transparency: capability to describe, inspect and reproduce the mechanisms through which AI systems make decisions

If Logic can provide explanations and justifications: Which Logic(s)?

Classical logic and extensions: infinitary logics, generalised quantifiers, epistemic, modal and temporal logics

Deviant logics: constructive logics, multi-valued logics, paraconsistent logics

Nonmonotonic logics: default logic, autoepistemic logic, defeasible logics, stable reasoning

What are the grounds for choice?

Internal principles of truth and inference: excluded middle, disjunctive syllogism, explosive axioms

General properties of inference and semantics: constructivity, computability, compactness, interpolation, cumulative inference, rationality

Expressive needs for applications: modal operators, special quantifiers, infinitary languages

Early days of Logics in AI: Preference for...

"Desirable" properties of inference:

cumulative, rational

$$\Pi \mathrel{|\!\sim} \varphi, \Pi \mathrel{|\!\sim} \psi \Rightarrow \Pi \cup \varphi \mathrel{|\!\sim} \psi$$

$$\Pi \mathrel{|\!\sim} \psi, \Pi \cup \varphi \not\mathrel{|\!\sim} \psi \Rightarrow \Pi \mathrel{|\!\sim} \neg\varphi$$

Computability: polynomial is better (but what about Datalog?)

Supraclassicality: add to classical logic rather than revise it (but remember Ptolomaic epicycles!)

Stable reasoning does not fare well

Not cumulative, not rational

Not polynomial

Not supraclassical (fails left and right absorption)

Oh Dear!!

Gelfond's adequacy criteria

1. *Clarity*: logical vocabulary should have a clear and intuitive meaning.
2. *Elegance*: the corresponding mathematics should be simple and elegant.
3. *Expressiveness*: the KR language should suggest systematic and elaboration tolerant representations of a broad class of phenomena of natural language, including belief, knowledge, defaults, causality and others.
4. *Relevance*: a large number of interesting computational problems should be reducible to reasoning about theories formulated in this language

Gelfond's adequacy criteria

But, good news! Gelfond rejects

- supraclassicality
- efficiency

Back to general principles of inference

Sometimes we may find a Lindström-style theorem, ie a property or properties that narrow down the class of logics to one or a small number

Lindström (1969): classical first-order logic is the strongest logic satisfying both:

- (countable) compactness: if a countable set of sentences has no model then some finite subset has no model

- (downard) Löwenheim-Skolem: if a sentence has an infinite model, it has a countable model

Back to general principles of inference

What happens when we extend first-order logic?

$L(Q_1)$ (“there exist at least $\aleph_1$ many”) is countably compact

$L_{\omega_1,\omega}$ satisfies the Löwenheim property

We have found Lindström-style properties for nonmonotonic logics

Under some general assumptions, equilibrium logic is the only system satisfying both

- atom definability

- well-supportedness

Even without uniqueness we may propose metatheoretic conditions on inference

consider the concepts of slide 1 in the context of epistemic logic programming and reasoning

- constraint monotonicity

- epistemic splitting

- foundedness

Metatheoretic conditions can be used in the task of **explication**

consists in transforming a more or less inexact concept into an exact one or, rather, in replacing the first by the second. We call the given concept the **explicandum**, *and the exact concept proposed to take the place of the first the* **explicatum**.

Strictly speaking , the question whether the solution [explicatum] is right or wrong makes no good sense because there is no clear-cut answer. The question should rather be whether the proposed solution is satisfactory (Carnap, 1950)

metatheoretic adequacy conditions for explication

Part of the task consists in specifying adequacy conditions that the explicatum should satisfy

A pre-formal analysis of the explicandum may suggest a series of properties desirable for the explicatum

In addition there are requirements for a satisfactory explication: similarity, exactness, fruitfulness, simplicity (compare with Gelfond)

exactness allows introducing the explication into a well-connected system of scientific concepts

another example

What may happen when we try to keep classical logic against all odds?

Consider the simple program rule $p \vee \neg p \rightarrow p$

On one approach this rule has the single intended model $\{p\}$. Why? Because $p \vee \neg p$ is a tautology

But the rule has no stable (equilibrium) model

So, what is a tautology?

Perhaps it is whatever we can add to a program without changing its stable models

In that case $p \vee \neg p$ is not a tautology; adding $p \vee \neg p$ to the program $p \rightarrow q; \neg p \rightarrow r$ (whose answer set is $\{r\}$) produces an additional answer set $\{p, q\}$.

And in this case we have a disjunctive program whose semantics is not in dispute. So, why regard as a tautology something that changes the meaning of a simple program?

It gets worse ...

Let Π be the propositional program:

$$\neg p \rightarrow p \tag{1}$$
$$\neg\neg p \rightarrow p \tag{2}$$

This has $\{p\}$ as its equilibrium or general stable model. Yet Shen, Wang, Eiter, Fink, Redl, Krennwallner & Deng (2014) say this suffers from a circular justification. Oh Dear!
But (1) is logically equivalent, even in constructive logic, to the formula $\neg\neg p$. So in Π p follows directly from (2) and re-written (1) by *modus ponens*! The inference to p is entirely monotonic and there is no issue of circular justification.

An alternative analysis

Since Π has the form $A \to C$ and $B \to C$, we should be able to infer that also $A \vee B \to C$. This holds as an axiom:

$$\vdash (A \to C) \wedge (B \to C) \to (A \vee B \to C) \tag{3}$$

in INT and even in minimal logic and in Anderson and Belnap's basic relevance logic **R**.
Applying to Π we should obtain

$$\neg p \vee \neg\neg p \to p \tag{4}$$

Since $\neg p \vee \neg\neg p$ is a tautology in classical logic as well as in **HT**, we should be able to infer p. Yet this is not the case, neither in FLP-semantics nor in the modified version. Since that accepts $\neg p \vee \neg\neg p$ as a tautology, the failure to infer p must be due to a failure to accept (3).

The same example in terms of rules

Think of Π as a set of rules

$$\frac{A \vee B \qquad \begin{array}{c}[A]\\ \vdots \\ C\end{array} \qquad \begin{array}{c}[B]\\ \vdots \\ C\end{array}}{C}$$

Figure: Rule of disjunction elimination.

So $\neg p \vee \neg\neg p \rightarrow p$ is derivable in constructive reasoning and the inference to p will follow in logics admitting the weak law of excluded middle. The approach of Shen et al lacks coherence because the type of logical reasoning that is permitted in determining when a rule atom is (non-circularly) inferable is quite different from the type of reasoning which would allow us to move from two different rules to a third one.

Summarising

Here we have a logic-based approach to KR

It rejects reasoning based on general stable models and equilibrium models, using spurious arguments about “circular justifications”. Confusing logical inference with justification or support.

It is not connected to any standard logical reasoning and it’s explicatum is not embedded into “a well-connected system of scientific concepts”. (Nor is it simple, intuitive, elegant, similar or fruitful!)

In the epistemic case it rejects metatheoretical adequacy conditions such as constraint monotonicity, splitting and foundedness in favour of pure intuition.

The authors are well-established and publish this work in leading AI journals.

Where to go from here?

In our project we consider adequacy conditions for logics

Internal adequacy conditions deal with the correctness of logical inferences, demostrated manually or automatically.

Logics may be enriched with additional features (causal graphs, argument trees, visualisations) to enhance their explanatory power.

External conditions refer to the rational acceptability of the logical system as a whole, and therefore the posterior acceptance of inference made and explanations given.

We may relates this to “social” or “cognitive” logics, but in my view the issue is not psychological but rather about acceptability by an ideally rational agent. (This may be analogous to questions about What is a mathematical proof?, ie possibly grounded in intuition but still “objective”.)

Presento

The design is clean

The rules are simple

The code is extensible

Open Source Fonts

This is Montserrat

This is Noto Sans

This is Lato (light)

This is inconsolata

This is Alegreya Sans small caps

Color Palette

BIG BOLD TEXT

RUN!