

AI Ethics & Philosophy of Logic



Antonis Kakas


University of Cyprus, Nicosia

&

Argument Theory, Paris

Moral Alignment

- **Extremely Difficult** problem
 - ~~What is morality?~~ ~~Which morality?~~
 - **How** to attain morality?

 - **How?** **Irrespective** of the what!
- 
- What is Logical structure of morality?**

Moral Alignment

How, irrespectively of the what!

Logical structure of morality

□ **Moral Norms vs Moral Guidelines**

□ **Moral Design vs Moral Development**

Moral Alignment

- **Known** problem from Human Society
 - There are **no "clever"** solutions.
- Follow the **old** solutions.
 - **Regulate** autonomous behaviour
- But **HOW?**

The Human Moral Alignment

Regulate Human Autonomous Behaviour
HOW?

- **Law, Regulations, Guidelines,**
- **Best Practices of Good & Bad, Nudging,**
 - **Encouragement, Fear of Punishment, Uncomfortable Feeling, ...**

Still fails! Extremely difficult problem!

The Human Moral Alignment

Regulate Human Autonomous Behaviour

HOW?

- **Nurturing & Coaching/Education**
- **Accountability via Explainability**

The How of Moral System Alignment

Regulate Autonomous Systems/Artifacts

- **FIRST PART:** Do what we do for any artifact that is released to the market.
 - A-priori approval via “Clinical” Tests
 - Akin to **Pharmaceutical Products**
 - Test Mental & Societal (moral) toxicity of AI

The How of Moral System Alignment

Regulate Autonomous Systems/Artifacts

■ SECOND PART: What "CLINICAL" TESTS?

- Task-oriented AI Systems (**not AI**)

- Efficacy & Ethicacy

■ Onus on the producer (c.f. **Pharma Case**)

PART 2

What is the Logical Structure of morality

The How of Moral Alignment

What is the **Logical Structure** of **morality**?

Moral Norms **Moral Guidelines**

Strict Logic/

Deontic Logic

???

Is there a **Logic** that can cover
the whole **spectrum**?

The How of Moral Alignment

What is the **Logical Structure** of morality?

Q: Is there a **Logic that can cover the whole **spectrum**?**

A: **Human Moral Reasoning?**

If so, this is **not strictly logical, nor it is **statistically** based.**

Enter Aristotle

The First Logician

Introduces: Strict Entailment
“Demonstration”

“Ἔστι δὴ συλλογισμὸς λόγος ἐν ᾧ τεθέντων τινῶν ἕτερόν τι τῶν κειμένων ἐξ ἀνάγκης συμβαίνει διὰ τῶν κειμένων.”

Aristotelian Syllogistic Logic

Enter Aristotle again!

The First Thinker of “Logic in Ethics”:

- **Logos over Pathos**
- **Practical Syllogisms for Ethical Actions**
- **“Scientific” Theory of Ethics”**

Enter Aristotle

The First Thinker of “Logic in Ethics”:

Aristotle recognizes:

- 1. The difficulty of the problem**
- 2. NOT a usual scientific theory.**
- 3. Strict Logic is NOT appropriate**

Aristotle:

The First Thinker of “Logic in Ethics”

The Nature of Moral Reasoning

“Moral reasoning is less rigorous than Scientific reasoning”

“To reason well we need to take into consideration both defeasible general principles and the normative relevant particular facts of the context and weigh the moral reasons they provide.”

“Leniency in reasoning [law] as a mechanism for the needed contextual correction in absolute reasoning.”

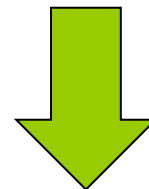
Aristotle: The First Thinker of “Logic in Ethics”

□ Deliberation

“It is a characteristic of men of **practical wisdom** to have **deliberated well.**”



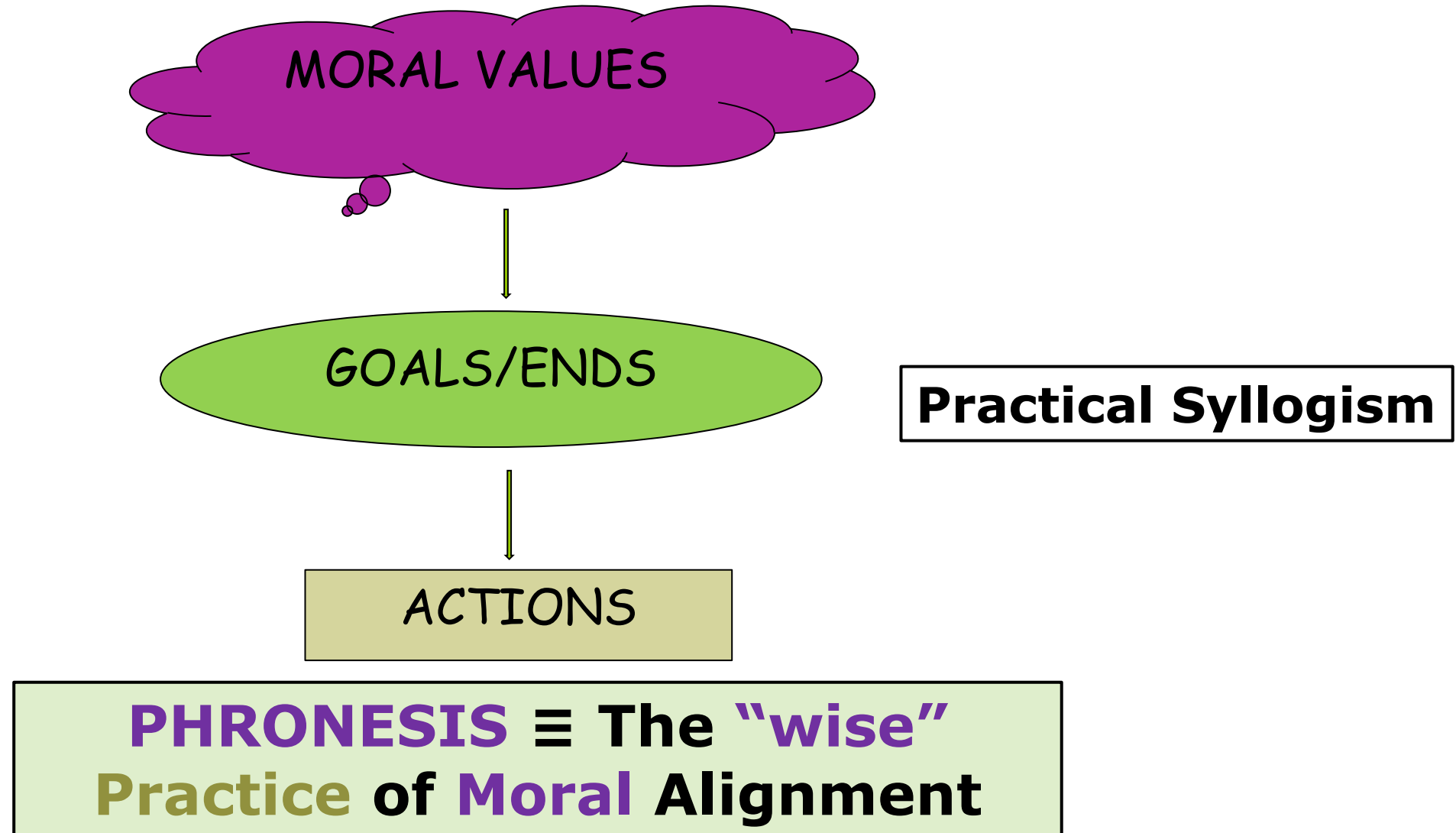
□ Phronesis (Prudence)



Computational Manifestation
“Concerned immediately with particulars”

□ Practical Syllogisms

Aristotle's System of Phronesis



Aristotle: The First Thinker of “Logic in Ethics”

What Logic?

- **Strict Logic for Practical Ethics? NO**
- **Ethics is non reducible to absolute moral principles**
- **Logic with “Premises” of “For the most part”**
 - **Logic with Leniency**

Back to

What is the **Logical Structure** of **morality**?

Moral Norms **Moral Guidelines**

Strict Logic/

Deontic Logic

???

Is there a **Logic** that can cover
the whole **spectrum**?

PART 3:
What is Logic?

Argumentation: Reasoning Universalis

Formal **Informal**

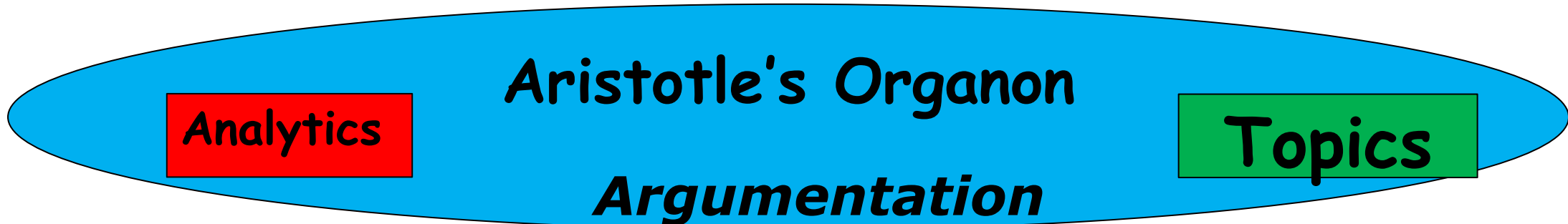
Flexibility of Argumentation



Intensity of Argumentation

Correct Thinking

Free Thinking



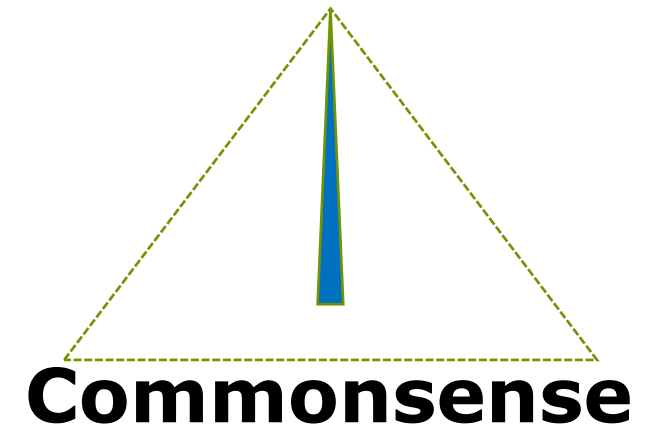
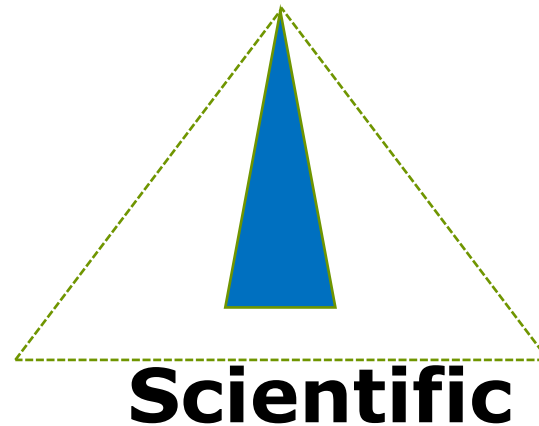
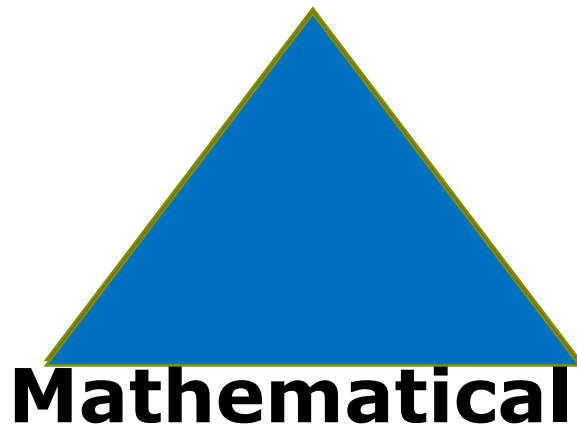
Dialectic **Argumentation**

4

Strict Rational Intelligence

VS

Natural Intelligence



Back to the Origin of Logic

" Ἔστι δὴ συλλογισμὸς **λόγος** ἐν ᾧ τεθέντων τινῶν ἕτερόν τι τῶν κειμένων **ἐξ ἀνάγκης** συμβαίνει διὰ τῶν κειμένων."

There is **no Logic** in Aristotle's writings!

Λόγος (Argument) – the basic notion/vehicle in **Λογική (Logic)**

λόγος = argument

ἐξ ἀνάγκης = necessarily

Formal Logical inference is the case of an absolute winner in the Argument Debate

What is **Logic**?

What makes an **Argument** **Logical**?

Aristotle: What is Logical Reasoning?

(Topics, Book A, 1st Sentence)

"The purpose of the present treatise is to discover a **method** by which we shall be able to **reason** from **generally accepted opinions** about any problem set before us and shall ourselves, when **sustaining** an **argument**, **avoid** saying anything **self-contradictory**."

Aristotle: Father of Argumentation

Normative Condition for Logical Reasoning:
Non-self-contradictory Argument

What is **Logic**?

What makes an **Argument Logical**?

Non-self-Contradictory

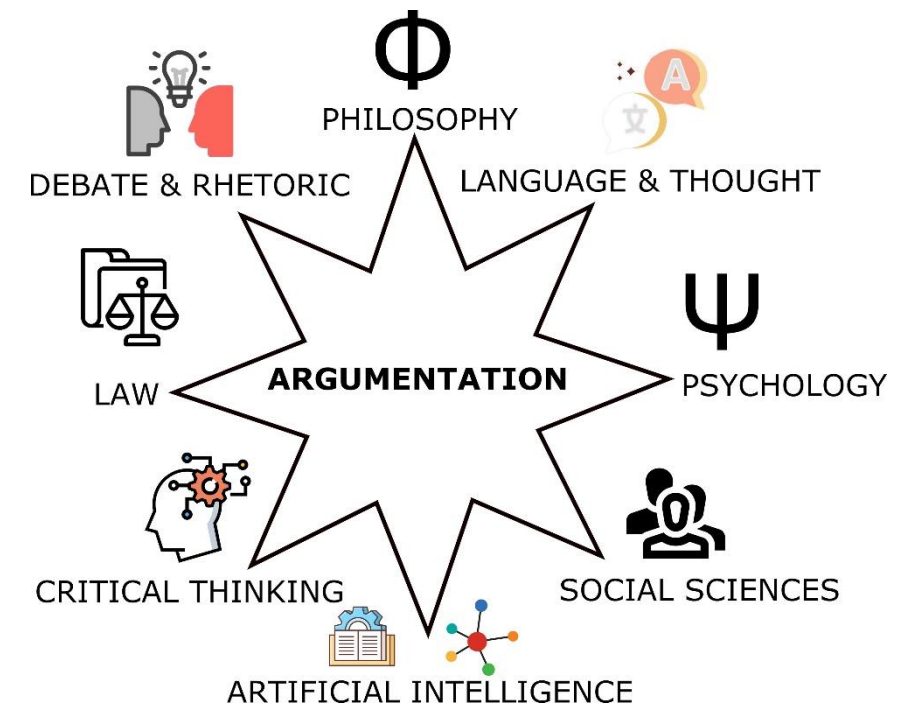
≈ **CL Satisfiable**

≈ **Cognitive Dissonance**

Cognitive support for Argumentation

Argumentation is native to Human Reasoning

- **Cognitive Psychology**
 - Mercier & Sperber:
 - “Why do humans reason?”
- **Behaviour Economics**
 - Thaler, Kaneman:
 - “Humans are not rational”
- **Computational Argumentation in AI models well Cognitive Empirical Data:**
 - **Syllogistic, Selection & Suppression Tasks**
- **Non-monotonic Logics in AI all reformulated in terms of Argumentation.**



Properties of Argumentation for Reasoning of AI Systems

- ***Flexible / Open***

- ***Contextual Reasoning***

- ***Cognitive***

- ***Compatibility with human thought-reasoning***

- ***Social***

- ***Channel of human-machine cooperation and synergy***

Argumentation Logic: Summary

The Logic of Natural Language

1. Reasoning with Guidelines

1. Equivalent to **Classical Logic** (At boundary)

2. Reasoning with Incompleteness & Inconsistency

1. No Logical paradoxes.

3. Explanatory Reasoning

1. Extracted from the **Argumentative Deliberation**

PART 4:
Explainable Logic – XAI & Morality?

Moral Alignment

HOW?

1. Reasoning with Normative Guidelines
2. **Accountability via Explainability**

But Why?

To Safeguard against Failure.

To allow Contestability, Debate and Change.

Moral Alignment

Via Explainability

- **Explanation justifies decision/action by connecting to moral values.**
- **Analysis of ethical dilemma & its resolution**
- **Seed for dialogue and argumentative debate.**

Supports a Continuous Process of building moral agents

Ethical Features of Explanations

- **Connects decision to moral values**
 - **Attributive:** adherence to moral value
 - **Contrastive:** non-adherence/violation of moral value
- **Reveals agent's beliefs of facts**
- **Internal Coherence: Non-dissonance**

An Ethical Clinical Test for any AI system

PART 5:
Moral Learning

Artificial Moral Learning

Can AI learn to be moral?

□ **Can AI develop Artificial Phronesis?**

□ **HOW: Moral Design vs Moral Development**

Aristotle: Only by “Habitual Learning”

Aristotle's Moral Learning

1. **Learn (defeasible) Ethical Principles.**
2. **Learn to Apply them.**

becoming Virtuous - Acquiring Phronesis

HOW to Learn?

- **A life-long endeavor**
- **Observe, deliberate and generalize**
- **From/Form: Virtuous Exemplars**
 - **Cultivate affection for good.**

Learning

- **Learning by nature is a defeasible process.**
 - **Target Language cannot be strict logic**
 - **Post-hoc Explanatory Models cannot be formulated in strict logic**

Argumentation Logic: The logic of Learning

Moral Machine Learning

Follow the **Natural Human** model:

- Learn from **Experience & Coaching**

- **Neural-Symbolic** approach
- **Moral Guidelines + Life-long adaptation/refinement**

- What is the appropriate training data?

- **“Digital life”** experience?

Moral Agency

1. What is a good or acceptable ethical operation of a system? Perfection or Sensitivity to special cases and adaptability.
2. How are ethical norms formulated within an AI system?
3. Strict Norm Compliance or Normative Guidance? Absolute Compliance vs Flexible Leniency
4. Ethics via Optimal Rationality or Dialectic Rationality of satisficing sustainable decisions
5. One shot Ethics by Rational Design or Habitual Ethicacy from continuous experience
6. What is an appropriate form of Ethical Data to train AI systems to Learn how to “live/operate” ethically.
7. How does explainability contribute to the Ethicacy of AI systems? What are good ethical characteristics of explanations?

Conclusion 0

Argumentation: **Reasoning Universalis**

Formal Deductive Reasoning

Human Reasoning in Natural Language

Conclusion 1

Argumentation

The Logical Foundation for Ethical Reasoning and Moral Learning

Conclusion 2

Argumentation

Covers the whole spectrum of Ethical Requirements: Norms to Guidelines.

Supports the Moral Vehicle of Explanation

Facilitates Moral Learning.

Conclusion 3

Aristotle on Computational Ethics

Not a case of formal reasoning

Defeasible/Lenient Logic

Not a one-shot process

Continuous Habitual Learning

Conclusion 4

Why this Tutorial?
(Why Philosophy?)

**AI is - needs to be -
outward looking.**

DEMO of COGNICA.

“At 5:45pm at ???”

Thanks

Let us close with a **joke.**

“Legal and Logical”