

# Systemic Reflections on GenAI and Legal Informatics

*Invited talk to the Last-JD (Law, Science and Technology Joint Doctorate)  
doctoral board*

Bologna, 14 April, 2026



Giovanni Sileno  
[g.sileno@uva.nl](mailto:g.sileno@uva.nl)

*University of Amsterdam*

An initial clarification...

An initial clarification...

What do we mean by Legal Informatics?

Legal informatics  $\neq$  Information law  
AI for Law  $\neq$  Law for AI

**Legal informatics  $\neq$  Information law**  
**AI for Law  $\neq$  Law for AI**

We devise and use computational  
methods for the study of and for  
applications in legal/normative  
systems at large

# **Legal informatics ≠ Information law**

## **AI for Law ≠ Law for AI**

We devise and use computational methods for the study of and for applications in legal/normative systems at large

We do legal (normative/policy) scholarship about computational processes and infrastructures

# Legal informatics ≠ Information law

## AI for Law ≠ Law for AI

We devise and use computational methods for the study of and for applications in legal/normative systems at large

We do legal (normative/policy) scholarship about computational processes and infrastructures

Here we **REGULATE**  
(contemporary, mainstream)  
technologies and their uses.

# Legal informatics $\neq$ Information law

## AI for Law $\neq$ Law for AI

We devise and use computational methods for the study of and for applications in legal/normative systems at large

Here we **DISCOVER/DESIGN** (new) techniques/processes for analysing or implementing regulations

We do legal (normative/policy) scholarship about computational processes and infrastructures

Here we **REGULATE** (contemporary, mainstream) technologies and their uses.

# The beginnings (?)

In order to compile and maintain a growing body of norms, Roman Law established a sophisticated **taxonomy** and **indexing systems**, transforming legal customs into a structured logical framework.

## C O R P U S J U R I S C I V I L I S,

P A N D E C T I S

Ad Florentinum archetypum expressis,  
INSTITUTIONIBUS, CODICE ET NOVELLIS,

*Addito textu Græco, ut & in Digestis & Codice, Legibus & Constitutionibus  
Græcis, cum optimis quibusque Editionibus collatis.*

*Cum notis integris, repetita quintum prælectionis,*

D I O N Y S I I G O T H O F R E D I, J C.

P R Æ T E R

JUSTINIANI EDICTA, LEONIS & ALIORUM IMPERATORUM NOVELLAS,  
Ac Canones Apostolorum, Græcè & Latinè, Feudorum Libros, Leges XII. Tabul. & alios ad jus  
pertinentes Tractatus, Fastos Consulares, Indicesque Titulorum ac Legum: & quæ-  
cunque in ultimis Parisienfi vel Lugdunenfi editionibus continentur,

*Huic editioni novè accefferunt*

Pauli receptæ Sententiæ cum selectis notis J. C U J A C I I

*Et sparsim ad universum Corpus*

ANTONII ANSELMO, A. F. A. N. J C. ANTVERP.

Observationes singulares, Remissiones & Notæ Juris Civilis, Canonici, & Novissimi  
ac in Praxi recepti differentiam continentes;

*Denique,*

Lectiones variz & Notæ selectæ Augustini, Bellonii, Goveani, Cujacii, Duareni, Ruffardi, Hottomanni,  
Contii, Roberti, Rævardi, Charondæ, Groti, Salmassii & aliorum.

*Opera & Studio*

SIMONIS VAN LEEUWEN, J C. LUGD. BAT.



# The beginnings (?)

In order to compile and maintain a growing body of norms, Roman Law established a sophisticated **taxonomy** and **indexing systems**, transforming legal customs into a structured logical framework.

This was the historical equivalent of **information retrieval** and **metadata standards**, treating *legal principles* as scalable protocols for the use of information.

## C O R P U S J U R I S C I V I L I S,

P A N D E C T I S

Ad Florentinum archetypum expressis,  
INSTITUTIONIBUS, CODICE ET NOVELLIS,

*Addito textu Græco, ut & in Digestis & Codice, Legibus & Constitutionibus  
Græcis, cum optimis quibusque Editionibus collatis.*

*Cum notis integris, repetita quintum prælectionis,*

D I O N Y S I I G O T H O F R E D I, J C.

P R Æ T E R

JUSTINIANI EDICTA, LEONIS & ALIORUM IMPERATORUM NOVELLAS,  
Ac Canones Apostolorum, Græcè & Latine, Feudorum Libros, Leges XII. Tabul. & alios ad jus  
pertinentes Tractatus, Fastos Consulares, Indicesque Titulorum ac Legum: & quæ-  
cunque in ultimis Parisienfi vel Lugdunenfi editionibus continentur,

*Huic editioni novè accefferunt*

Pauli receptæ Sententiæ cum selectis notis J. C U J A C I I

*Et sparsim ad universum Corpus*

A N T O N I I A N S E L M O, A. F. A. N. J C. A N T V V E R P.

Observationes singulares, Remissiones & Notæ Juris Civilis, Canonici, & Novissimi  
ac in Praxi recepti differentiam continentes,

*Denique,*

Lectiones variz & Notæ selectæ Augustini, Bellonii, Goveani, Cujacii, Duareni, Ruffardi, Hottomanni,  
Contii, Roberti, Revardi, Charondæ, Grotii, Salmasii & aliorum.

*Opera & Studio*

S I M O N I S V A N L E E U W E N, J C. L U G D. B A T.



## The beginnings (?)

Leibniz (also a lawyer) strongly believed that calculation would be the key to settle ***legal disputes***, all human conflicts and disagreements.



Leibniz: ***Calculemus!*** ~1686

## The beginnings (?)

Leibniz (also a lawyer) strongly believed that calculation would be the key to settle **legal disputes**, all human conflicts and disagreements

**...using characteristics numbers instead of concepts:**

animal = 2

rational = 3

man = rational animal =  $2 \times 3 = 6$



Leibniz: ***Calculemus!*** ~1686

# The beginnings (?)

...building upon and further  
strengthening the analytical tradition!

|                     |                |   |             |
|---------------------|----------------|---|-------------|
| iustum, licitum     | } est quicquid | { | possibile   |
| injustum, illicitum |                |   | impossibile |
| æquum, debitum      |                |   | necessarium |
| indebitum           |                |   | omissibile  |

indifferens est simul iustum et omissibile



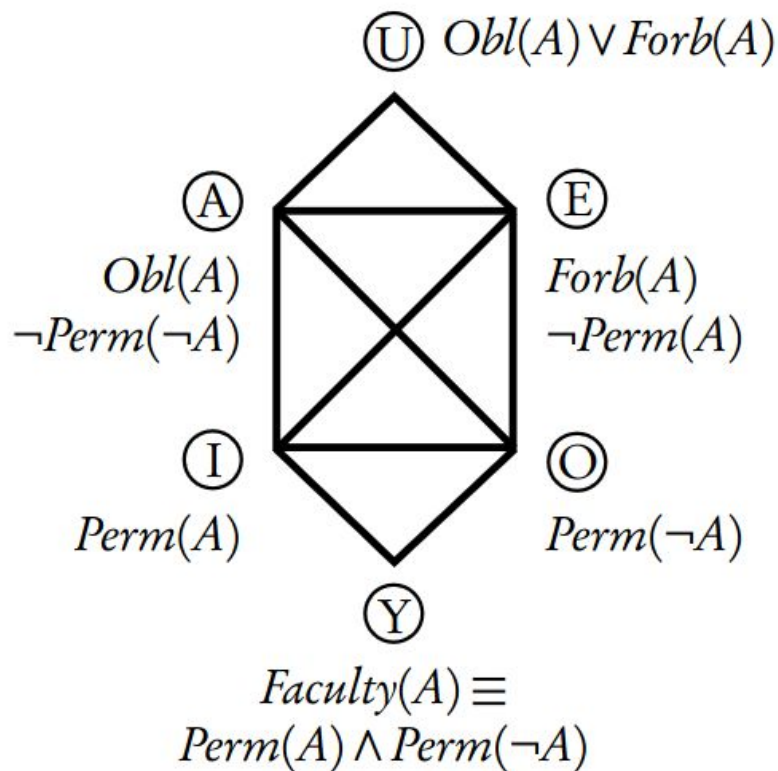
Leibniz: ***Calculus!*** ~1686

# The beginnings (?)

...building upon and further  
strengthening the analytical tradition!

|                     |                  |             |
|---------------------|------------------|-------------|
| iustum, licitum     | } est quicquid { | possibile   |
| injustum, illicitum |                  | impossibile |
| æquum, debitum      |                  | necessarium |
| indebitum           |                  | omissibile  |

indifferens est simul iustum et omissibile



**deontic logic**, core element of theories of normative systems

# A "fresh" a-new restart

L. Thorne McCarty

Reflections on TAXMAN: An Experiment in Artificial  
Intelligence and Legal Reasoning\*

## Table of Contents

|                                                                |                           |
|----------------------------------------------------------------|---------------------------|
| I. Introduction                                                | A. States and Events      |
| II. Some Fundamentals of Corporate Re-<br>organization Law     | B. Description Mechanisms |
| III. Some Fundamentals of Semantic in-<br>formation Processing | C. Analysis Mechanisms    |
| IV. Implementations                                            | V. Results                |
|                                                                | VI. Feasible Extensions   |
|                                                                | VII. Conclusion           |

## I. Introduction

This article addresses – in a highly unorthodox way – a series of questions which are surely among the most orthodox that can be found anywhere in the legal and jurisprudential literature. The questions are concerned primarily with legal concepts: What is the structure of a legal concept? What is the process by which legal concepts are transformed and modified? What are the arguments that lawyers characteristically make with respect to legal concepts? These are the kinds of questions that the Analytical Jurists once treated as essential, and the Realists dismissed as irrelevant; the kinds of questions that Cardozo ascribed to the Method of Philosophy in the judicial process, as opposed to the Method of Sociology<sup>1</sup>. In short, I will be raising here some of the classical problems of legal reasoning, in their most classical form.

McCarty, **TAXMAN** – 1972-77

# A “fresh” a-new restart

***Generic logic solvers  
as Prolog ⇒***

L. Thorne McCarty

Reflections on TAXMAN: An Experiment in Artificial  
Intelligence and Legal Reasoning\*

## Table of Contents

|                                                                |                           |
|----------------------------------------------------------------|---------------------------|
| I. Introduction                                                | A. States and Events      |
| II. Some Fundamentals of Corporate Re-<br>organization Law     | B. Description Mechanisms |
| III. Some Fundamentals of Semantic In-<br>formation Processing | C. Analysis Mechanisms    |
| IV. Implementations                                            | V. Results                |
|                                                                | VI. Feasible Extensions   |
|                                                                | VII. Conclusion           |

## I. Introduction

This article addresses – in a highly unorthodox way – a series of questions which are surely among the most orthodox that can be found anywhere in the legal and jurisprudential literature. The questions are concerned primarily with legal concepts: What is the structure of a legal concept? What is the process by which legal concepts are transformed and modified? What are the arguments that lawyers characteristically make with respect to legal concepts? These are the kinds of questions that the Analytical Jurists once treated as essential, and the Realists dismissed as irrelevant; the kinds of questions that Cardozo ascribed to the Method of Philosophy in the judicial process, as opposed to the Method of Sociology<sup>1</sup>. In short, I will be raising here some of the classical problems of legal reasoning, in their most classical form.

McCarty, TAXMAN – 1972-77

# A “fresh” a-new restart

***Generic logic solvers  
as Prolog ⇒***

L. Thorne McCarthy

Reflections on TAXMAN: An Experiment in Artificial  
Intelligence and Legal Reasoning\*

## Table of Contents

I. Introduction  
II. Some Fundamentals of Corporate Re-  
organization Law  
III. Some Fundamentals of Semantic In-  
formation Processing  
IV. Implementations

A. States and Events  
B. Description Mechanisms  
C. Analysis Mechanisms  
V. Results  
VI. Feasible Extensions  
VII. Conclusion

## I. Introduction

This article addresses – in a highly unorthodox way – a series of questions which are surely among the most orthodox that can be found anywhere in the legal and jurisprudential literature. The questions are concerned primarily with legal concepts: What is the structure of a legal concept? What is the process by which legal concepts are transformed and modified? What are the arguments that lawyers characteristically make with respect to legal concepts? These are the kinds of questions that the Analytical Jurists once treated as essential, and the Realists dismissed as irrelevant; the kinds of questions that Cardozo ascribed to the Method of Philosophy in the judicial process, as opposed to the Method of Sociology<sup>1</sup>. In short, I will be raising here some of the classical problems of legal reasoning, in their most classical form.

McCarthy, TAXMAN – 1972-77

## ARTICLES

# THE BRITISH NATIONALITY ACT AS A LOGIC PROGRAM

*The formalization of legislation and the development of computer systems to assist with legal problem solving provide a rich domain for developing and testing artificial-intelligence technology.*

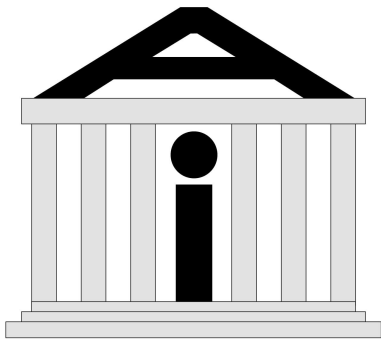
**M. J. SERGOT, F. SADRI, R. A. KOWALSKI, F. KRIWACZEK, P. HAMMOND, and H. T. CORY**

There are essentially two kinds of law, *case law*, determined by earlier court decisions, and *statutes*, determined by legislation. Substantial amounts of statutory law are basically definitional in nature and

the accuracy of the implementation and makes the implementation easier to maintain as the legislation changes and as case law evolves to augment the legislation.

Sergot, Sadri, Kowalski et al. – 1986

# In the heyday of AI & Law: 80-90s



International Association for Artificial  
Intelligence and Law (IAAIL)  
since 1987 => ICAIL conference

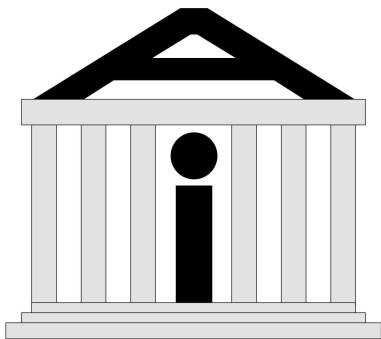


# JURIX

The Foundation for Legal Knowledge Based Systems

Foundation for Legal Knowledge  
Based Systems (JURIX)  
since 1988 => JURIX conference

# In the heyday of AI & Law: 80-90s



International Association for Artificial Intelligence and Law (IAAIL)

since 1987 => ICAIL conference

ICAIL2026 @ Singapore

<https://site.smu.edu.sg/icail-2026>

coming soon!



# JURIX

The Foundation for Legal Knowledge Based Systems

Foundation for Legal Knowledge Based Systems (JURIX)

since 1988 => JURIX conference

JURIX2026 @ Toulouse

call for papers in May/June 2026,

submission in September 2026

conference in December 2026

# Shallow-Knowledge vs Deep-Knowledge approaches

information  
retrieval tools

natural languages  
processing techniques

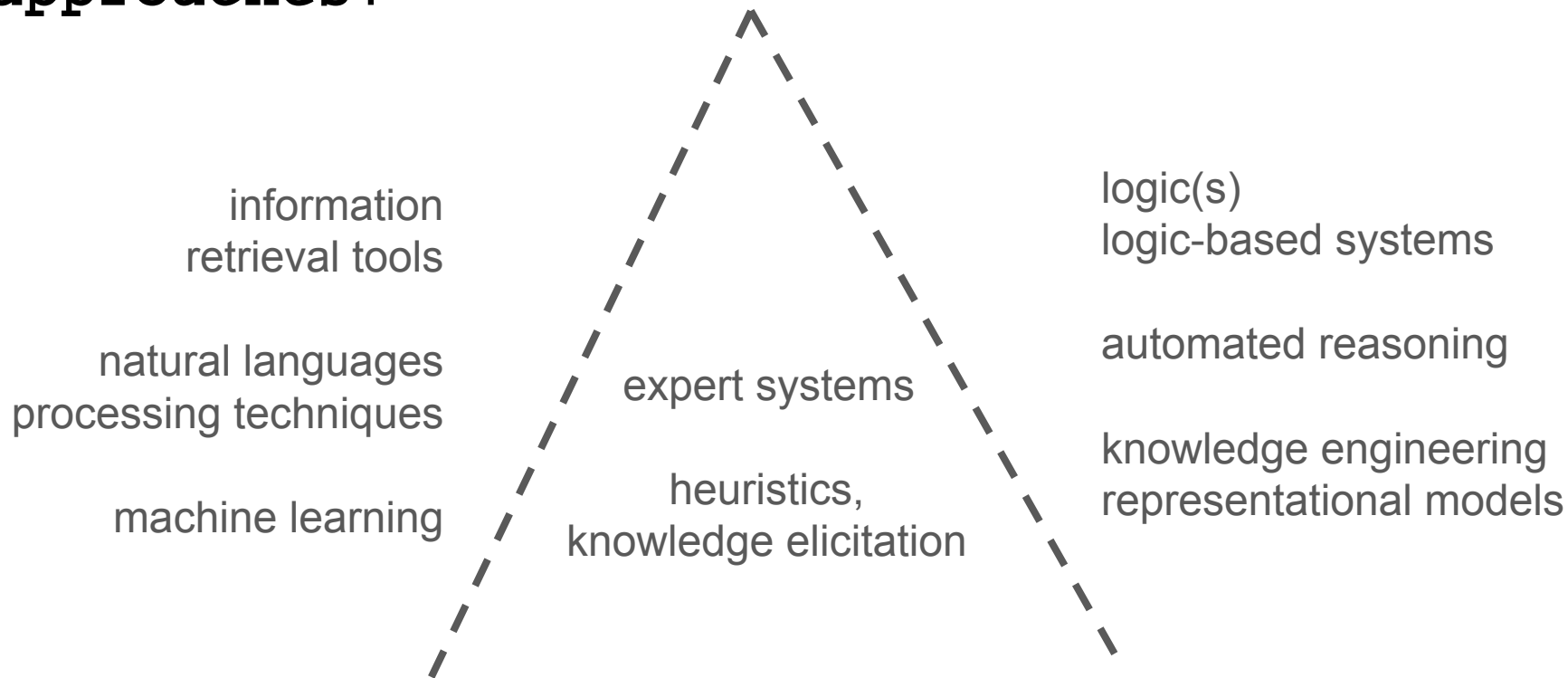
machine learning

logic(s)  
logic-based systems

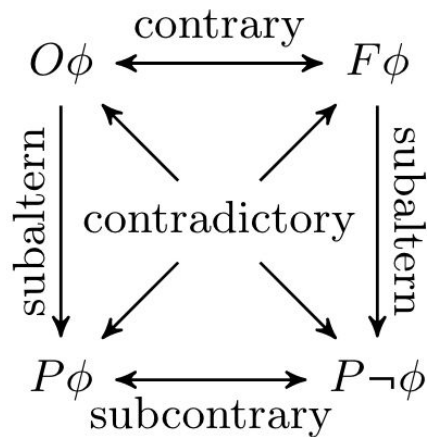
automated reasoning

knowledge engineering  
representational models

# Shallow-Knowledge vs Deep-Knowledge approaches+



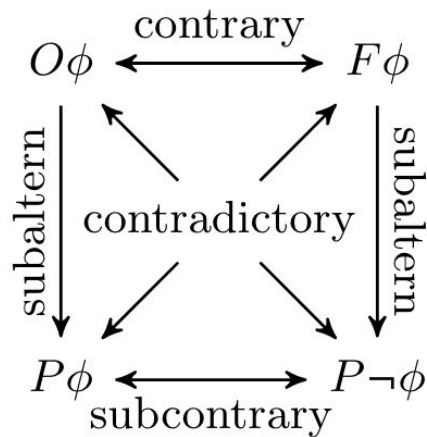
# Deontic logics and Normative systems



Deontic logics

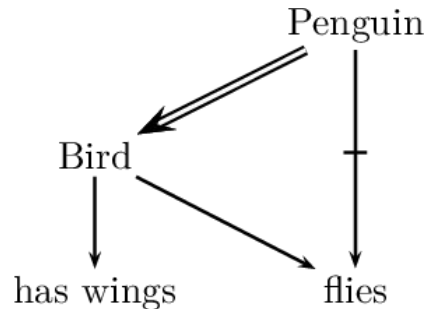
relevant concepts: **normative modalities, violations,**  
...

# Deontic logics and Normative systems



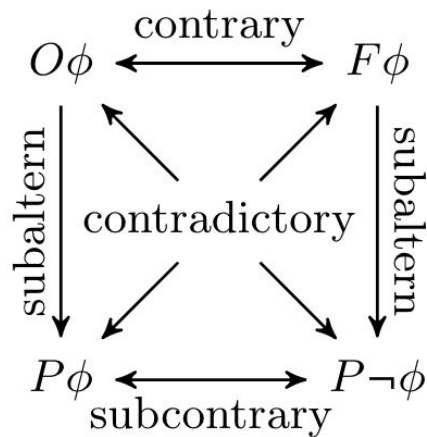
Deontic logics

relevant concepts: **normative modalities, violations, conflicts, priorities**



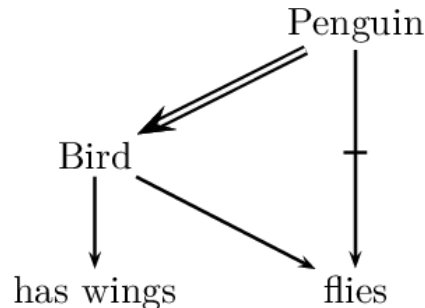
non-monotonic (defeasible) aspects

# Deontic logics and Normative systems



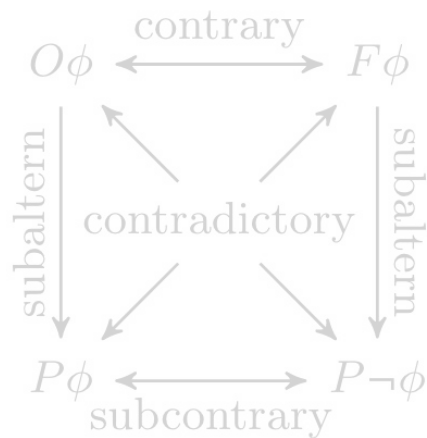
Deontic logics

relevant concepts: **normative modalities**,  
**violations, conflicts, priorities** + **constitutive aspects?**

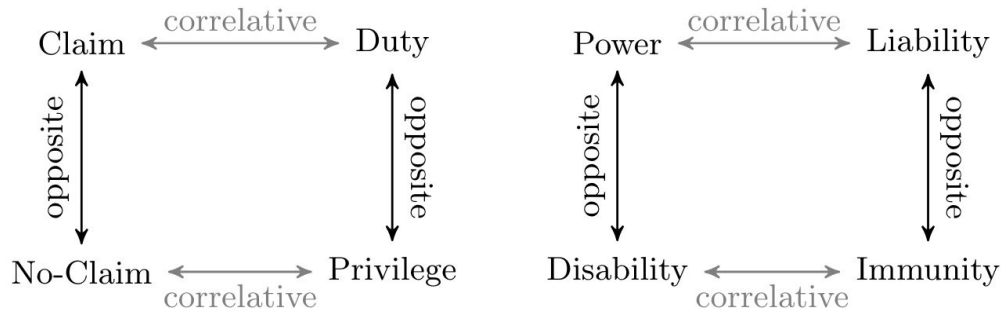


non-monotonic (defeasible) aspects

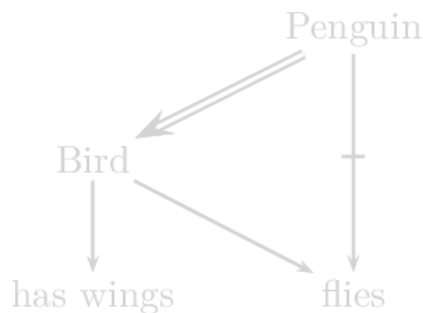
# Deontic logics and Normative systems



Deontic logics



Hohfeld's framework of normative concepts



relevant concepts:  
**power, roles,**  
**norm dynamics**

non-monotonic (defeasible) aspects

# Case-based reasoning

113

*Artificial Intelligence and Law 1: 113–208, 1992.*  
© 1992 Kluwer Academic Publishers. Printed in the Netherlands.

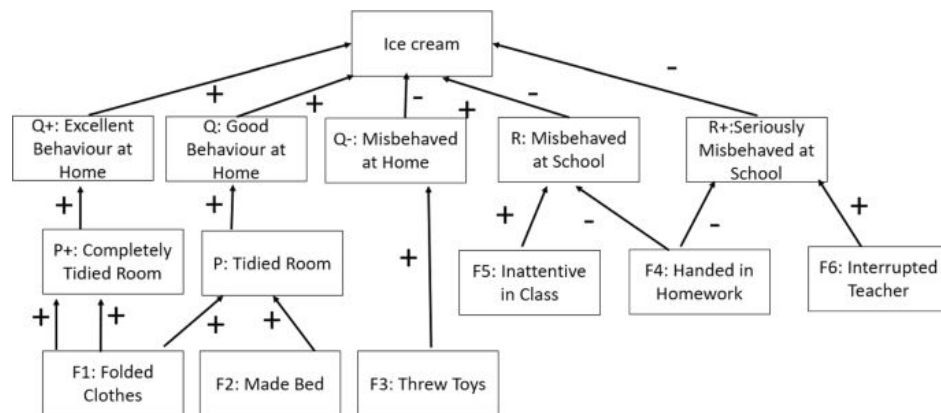
## Case-Based Reasoning and its Implications for Legal Expert Systems

KEVIN D. ASHLEY

*Assistant Professor of Law and Intelligent Systems,  
University of Pittsburgh,  
School of Law  
and  
Learning Research and Development Center,  
Pittsburgh, PA 15260, U.S.A.  
e-mail: ashley@vms.cis.pitt.edu  
tel: (412) 648-1495, 624-7496*

(Received 16 November 1991, accepted 8 October 1992)

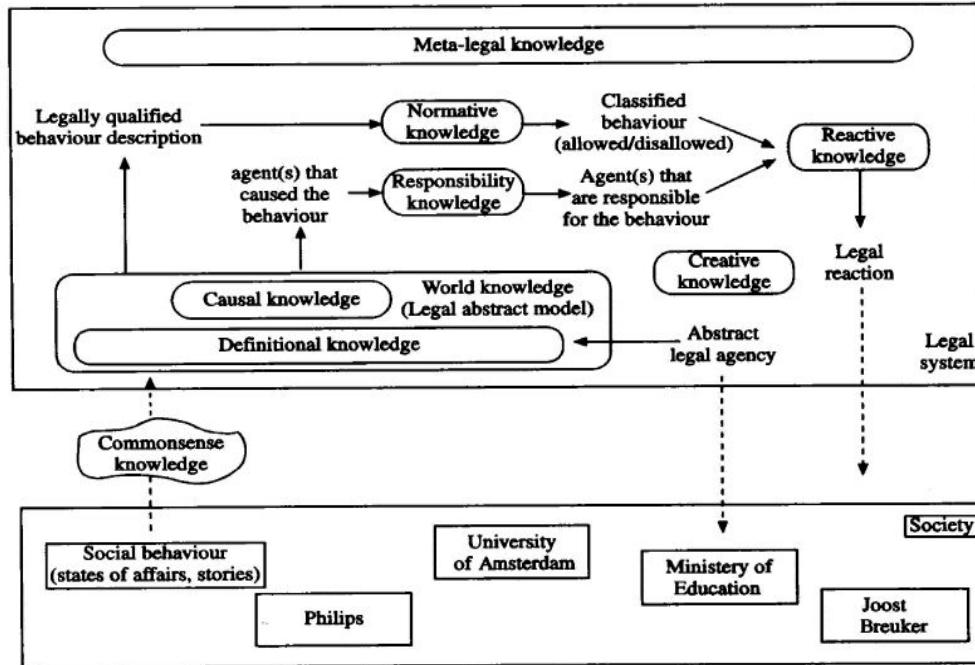
**Abstract.** Reasoners compare problems to prior cases to draw conclusions about a problem and guide decision making. All Case-Based Reasoning (CBR) employs some methods for generalizing from cases to support indexing and relevance assessment and evidences two basic inference methods: constraining search by tracing a solution from a past case or evaluating a case by comparing it to past cases. Across domains and tasks, however, humans reason with cases in subtly different ways evidencing different mixes of and mechanisms for these components.



relevant concepts: **factors** and **dimensions**

Ashley, Bench-Capon, Horty, ...

# Expert Systems

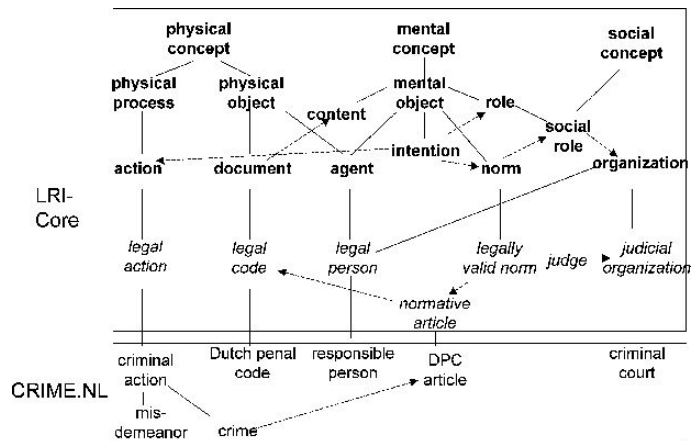


relevant aspects:

- knowledge domains, coupled with
- problem-solving methods and heuristics per type of task

Breuker, Valente, ...

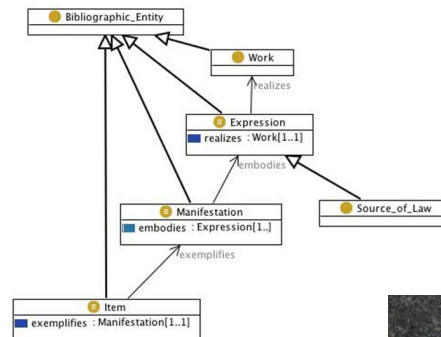
# Ontologies, and the Semantic Web



LRI-core (legal core ontology), LKIF-core  
– applied eg. to improve search

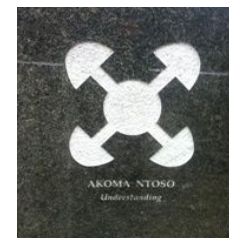
PrOnto: Privacy Ontology  
for Legal Compliance

MeLON: Method-ology for  
building Legal Ontology



CEN MetaLEX  
for legal documents

LegalRuleML, rule interchange  
language for the legal domain.



Akoma Ntoso

Breuker, Winkels, Van Engers, Hoekstra, Boer, ...  
Palmirani, Casanovas, Sartor, Francesconi...

# Recommender Systems

## Towards a Legal Recommender System

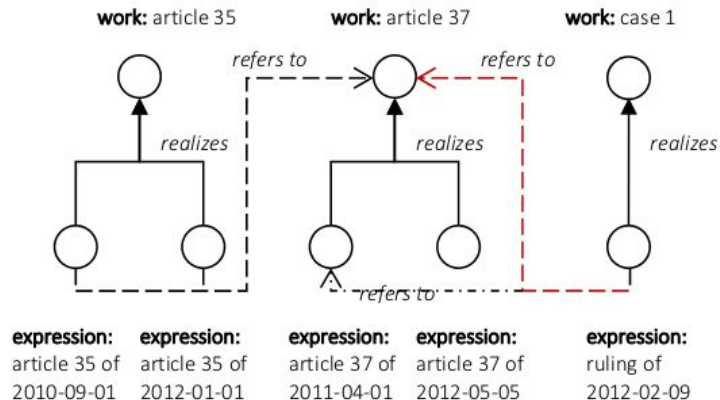
Radboud WINKELS<sup>1</sup>, Alexander BOER,  
Bart VREDEBREGT & Alexander van SOMEREN  
*Leibniz Center for Law, University of Amsterdam*

**Abstract.** In this paper we present the results of ongoing research aimed at a legal recommender system where users of a legislative portal receive suggestions of other relevant sources of law, given a focus document. We describe how we make references in case law to legislation explicit and machine readable, and how we use this information to adapt the suggestions of other relevant sources of law. We also describe an experiment in categorizing the references in case law, both by human experts and unsupervised machine learning. Results are tested in a prototype for Immigration Law.

**Keywords.** Semantic Web, citations, MetaLex, network analysis, cluster analysis, reasons for citing

### 1. Introduction

More and more sources of law are freely available online in the Netherlands, but also in the rest of Europe and the world. Most of the time however, these are stand-alone databases, containing one type of documents, not linked to other sources. For instance the Dutch portal for case law – rechtspraak.nl – contains a (small) part of all judicial decisions in the Netherlands. Case citations in these decisions are sometimes explicitly linked, references to legislation are not.<sup>2</sup>



network analysis in law

Winkels, ...

# Recommender Systems & NLP methods

## Towards a Legal Recommender System

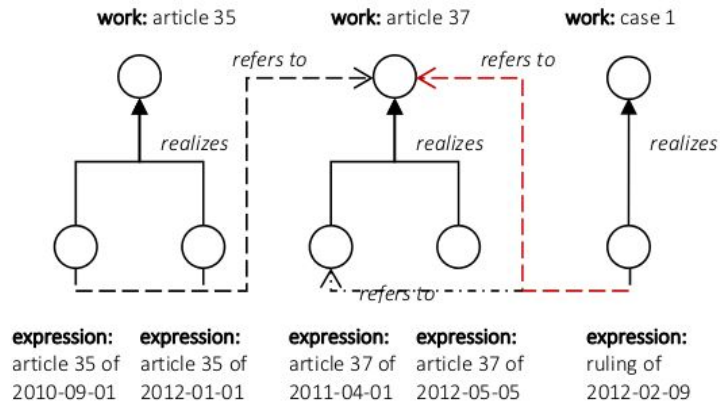
Radboud WINKELS<sup>1</sup>, Alexander BOER,  
Bart VREDEBREGT & Alexander van SOMEREN  
*Leibniz Center for Law, University of Amsterdam*

**Abstract.** In this paper we present the results of ongoing research aimed at a legal recommender system where users of a legislative portal receive suggestions of other relevant sources of law, given a focus document. We describe how we make references in case law to legislation explicit and machine readable, and how we use this information to adapt the suggestions of other relevant sources of law. We also describe an experiment in categorizing the references in case law, both by human experts and unsupervised machine learning. Results are tested in a prototype for Immigration Law.

**Keywords.** Semantic Web, citations, MetaLex, network analysis, cluster analysis, reasons for citing

### 1. Introduction

More and more sources of law are freely available online in the Netherlands, but also in the rest of Europe and the world. Most of the time however, these are stand-alone databases, containing one type of documents, not linked to other sources. For instance the Dutch portal for case law – rechtspraak.nl – contains a (small) part of all judicial decisions in the Netherlands. Case citations in these decisions are sometimes explicitly linked, references to legislation are not.<sup>2</sup>



network analysis in law  
topic modelling (possibly connecting to factors)  
argument mining...

Winkels, Ashley, Atkinson, Wyner,  
and many others...

# Formal argumentation

Arg1  $\xrightarrow{\text{attacks}}$  Arg2

Dung's framework

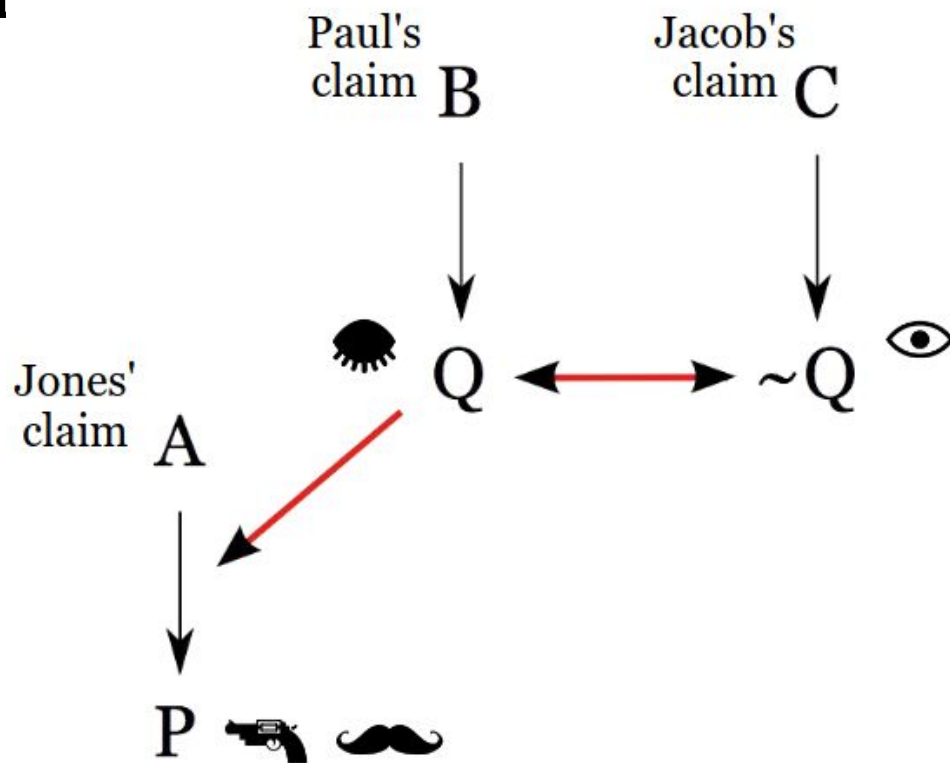
arguments and attack relationships

# Formal argumentation

Arg1  $\xrightarrow{\text{attacks}}$  Arg2

Dung's framework

arguments and attack relationships

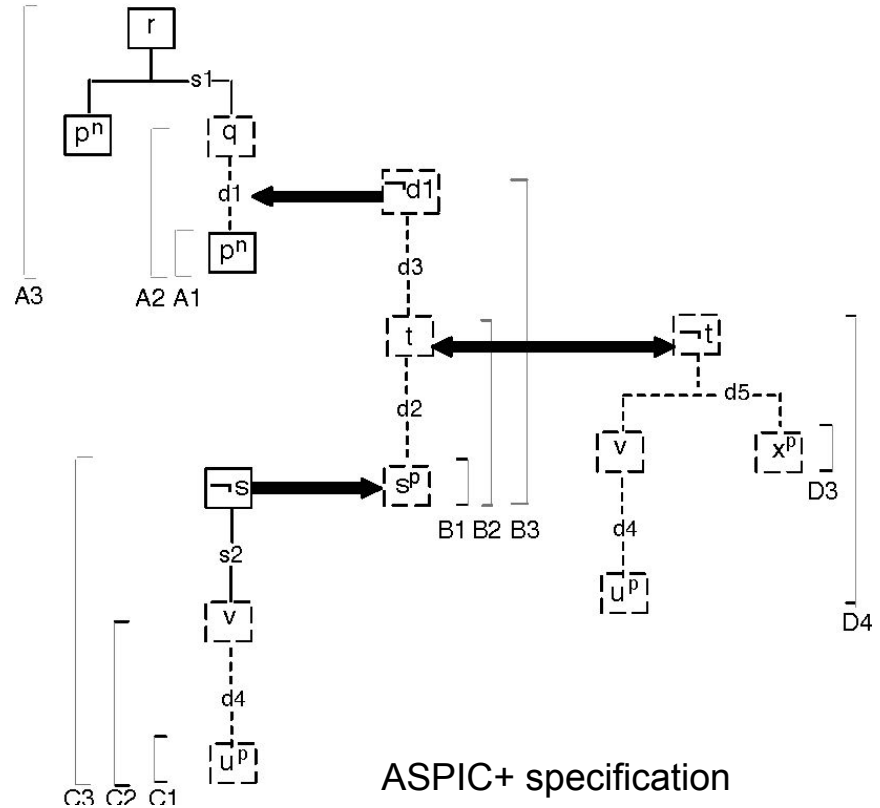


# Formal argumentation

$\text{Arg1} \xrightarrow{\text{attacks}} \text{Arg2}$

Dung's framework

arguments and attack relationships

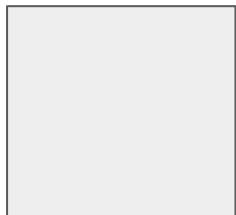


**Now? Use of LLMs is exploding everywhere!**

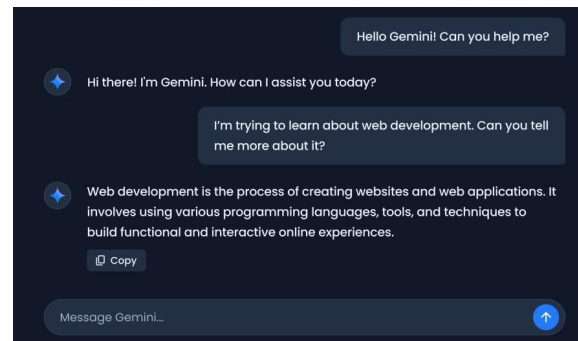
- The underlying assumption is essentially the same as Leibniz: reasoning as crunching numbers (although on latent spaces)...

# Typical use of LLMs

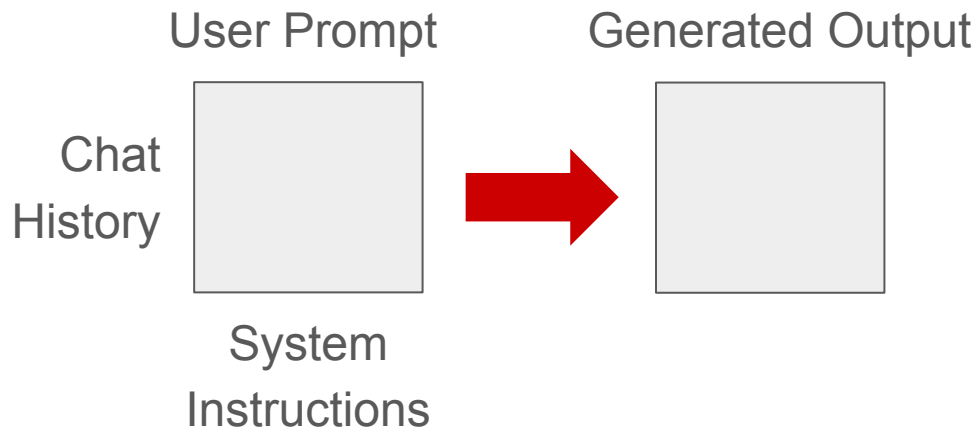
Input Prompt



Generated Output



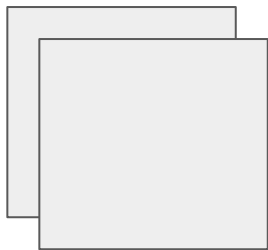
# Typical use of LLMs



# Typical use of LLMs for coding



Code



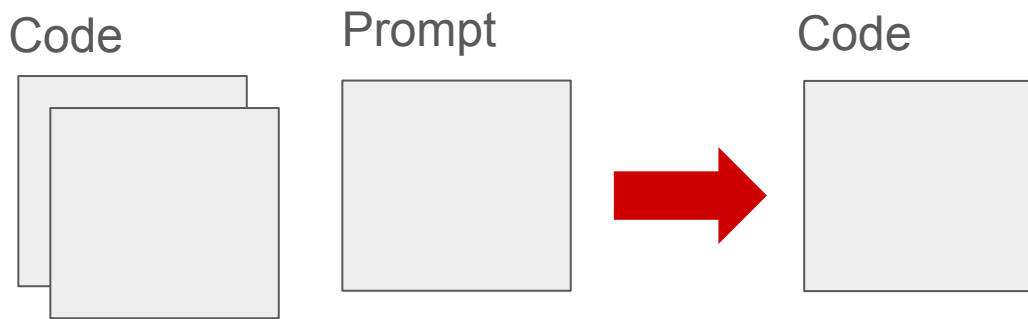
Prompt



Code

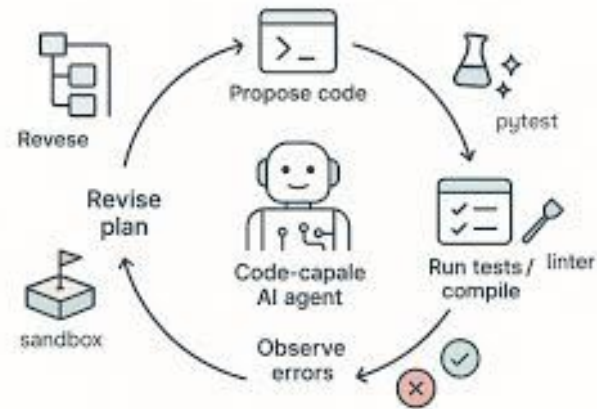


# Typical use of LLMs for coding

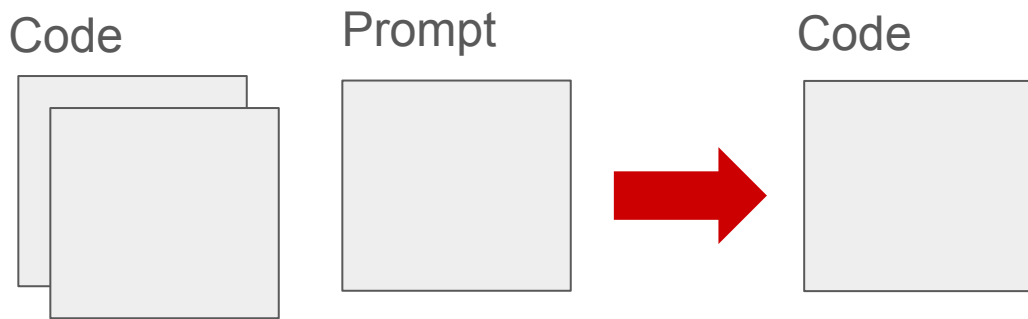


with Agentic AI:

- text for conceiving/designing the program,
- code for developing the program and
- code for verifying, validating, and deploying the program...



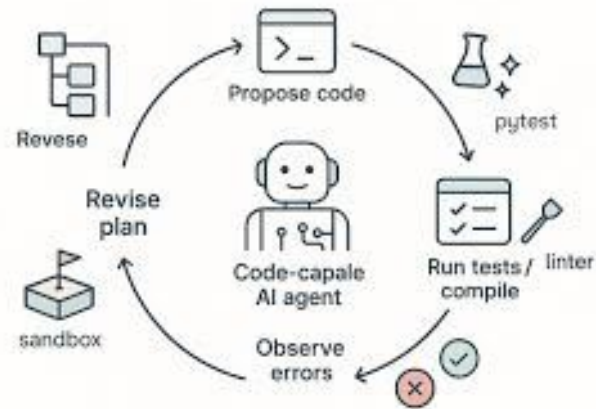
# Typical use of LLMs for coding



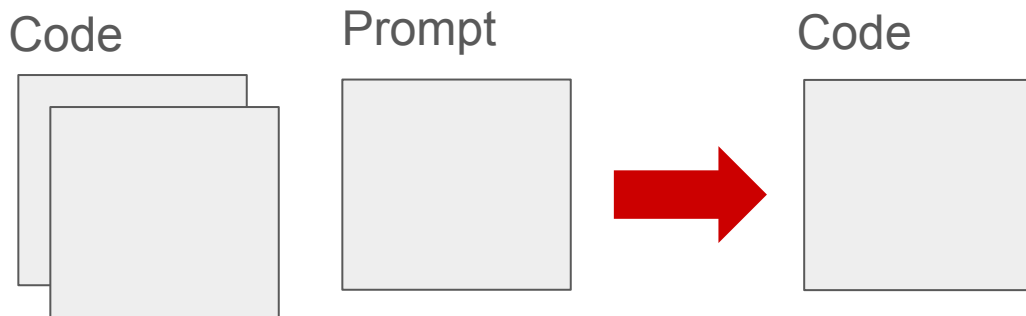
**where's the human?**

with Agentic AI:

- text for conceiving/designing the program,
- code for developing the program and
- code for verifying, validating, and deploying the program...



# Typical use of LLMs for coding



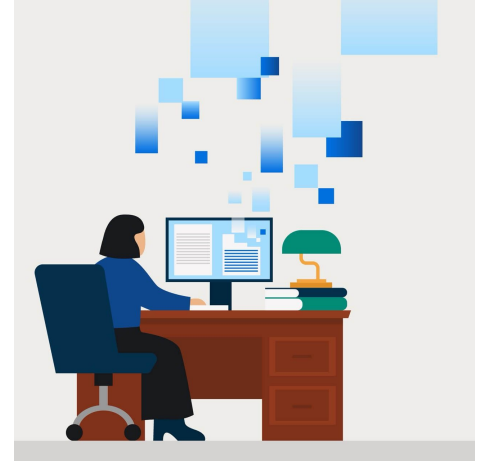
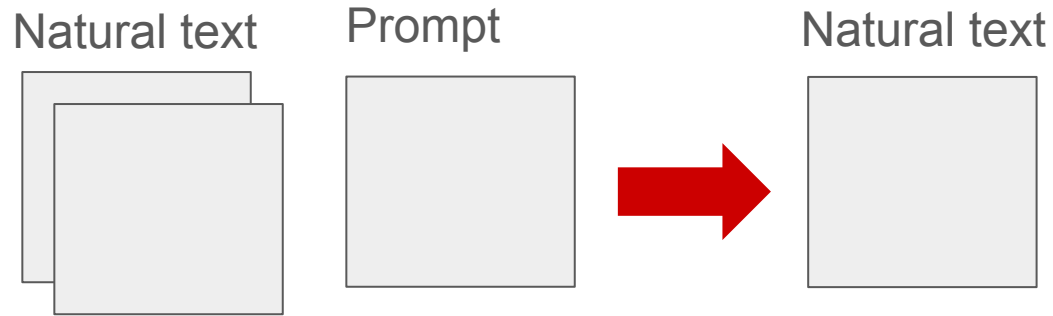
←  [r/learnprogramming](#) · 3mo ago  
Low-Tune-1869

**CS student here.. no one I know actually writes code anymore. We all use AI. Is this just how it is now?**

I'm a CS student, and at this point AI does all the coding. Not most of it. All of it. My classmates and I don't write code anymore. We describe the problem, get a full solution/help from AI, and then our job is to understand what the AI produced.

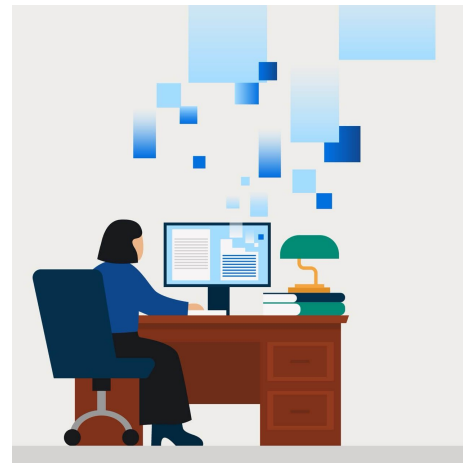
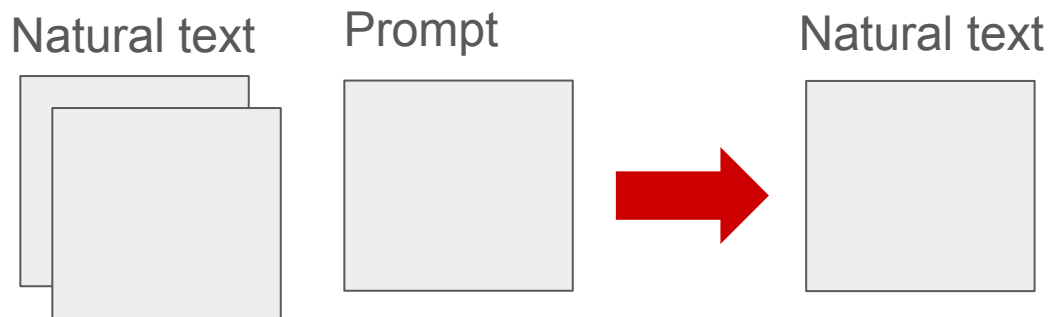
We read the code, follow the logic, but the solution itself is entirely generated. Writing code line by line just doesn't happen. I don't think anyone can write a method that return something in my class without ai

# Typical use of LLMs in legal practice



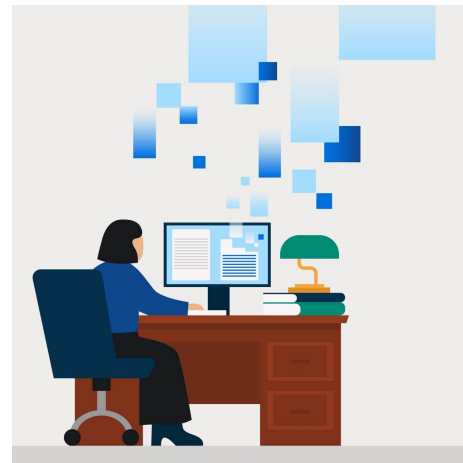
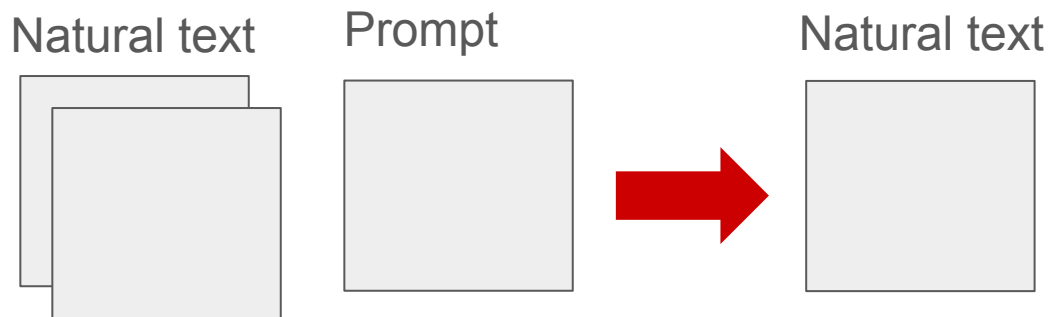
plenty of legal tasks are based on texts.

# Typical use of LLMs in legal practice



plenty of legal tasks are based on texts.  
It seems LLMs offer a “general-purpose” instrument to  
to perform almost all legal tasks...

# Typical use of LLMs in legal practice

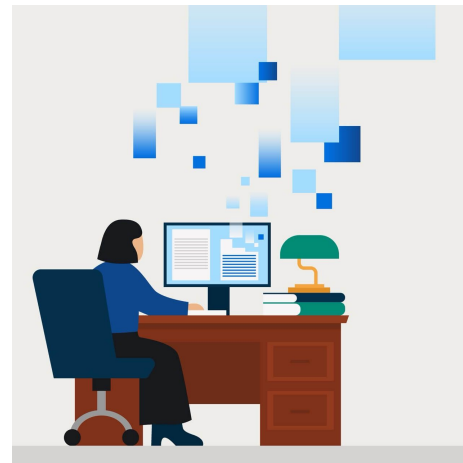
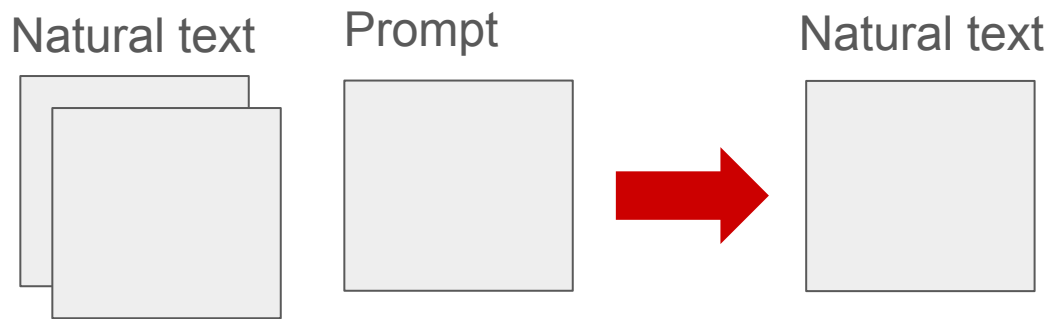


plenty of legal tasks are based on texts.  
It seems LLMs offer a “general-purpose” instrument to  
to perform almost all legal tasks...



*should we throw away  
all what done so far in  
**legal informatics?***

# Typical use of LLMs in legal practice



Well, things can still go (very) wrong...

Sept 03, 2025

**Lawyer caught using AI-generated false citations in court case penalised in Australian first**

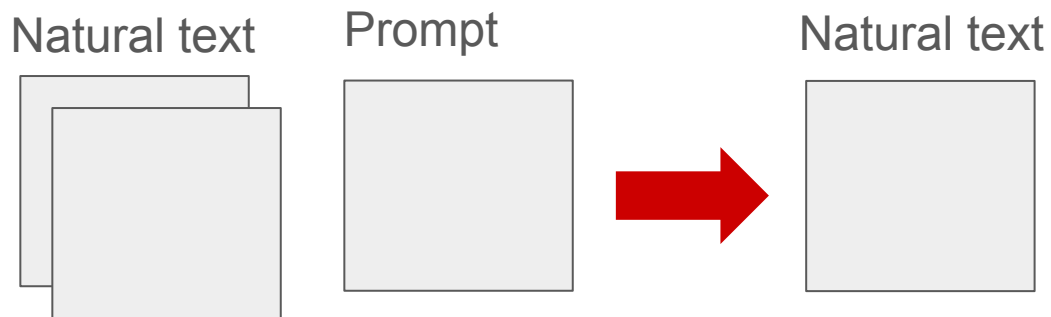
Victorian solicitor stripped of his ability to practise as a principal lawyer after acknowledging he did not verify case list

Apr 06, 2026

**Mata v. Avianca, Inc. (Federal Court, New York)**

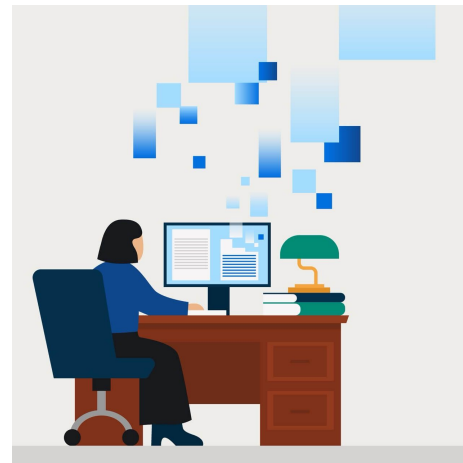
The Mata v. Avianca, Inc. case involves the plaintiff's lawyers who used AI to draft a motion that included fabricated citations, which misled the court. Opposing counsel identified false information in the filings, leading to hearings and sanctions, with the court emphasizing the importance of accurate legal documentation.

# Typical use of LLMs in legal practice

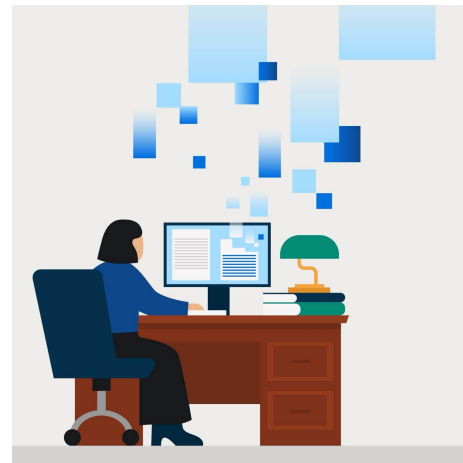
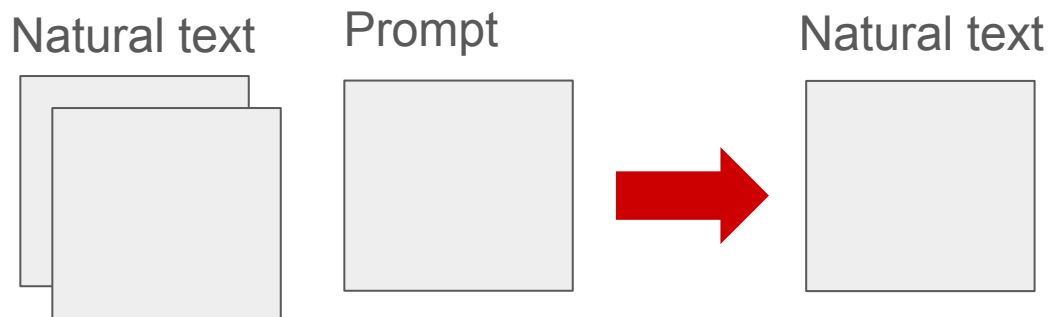


Well, things can still  
go (very) wrong...

*and we do not know easily WHY...*



# Typical use of LLMs in legal practice



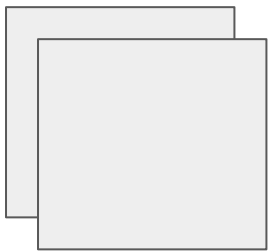
Well, things can still  
go (very) wrong...

*and we do not know easily WHY...*

“allucitation” is a **feature**, not an error!

# Why does it seem to work so well?

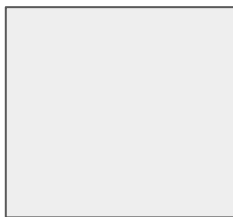
Natural text



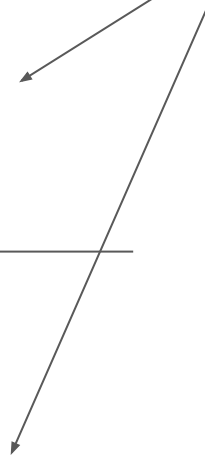
Prompt



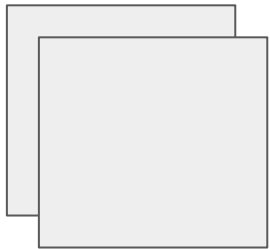
Natural text



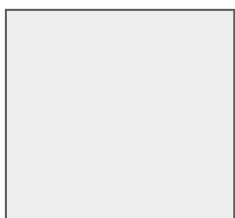
Domain of *discourse*



Code



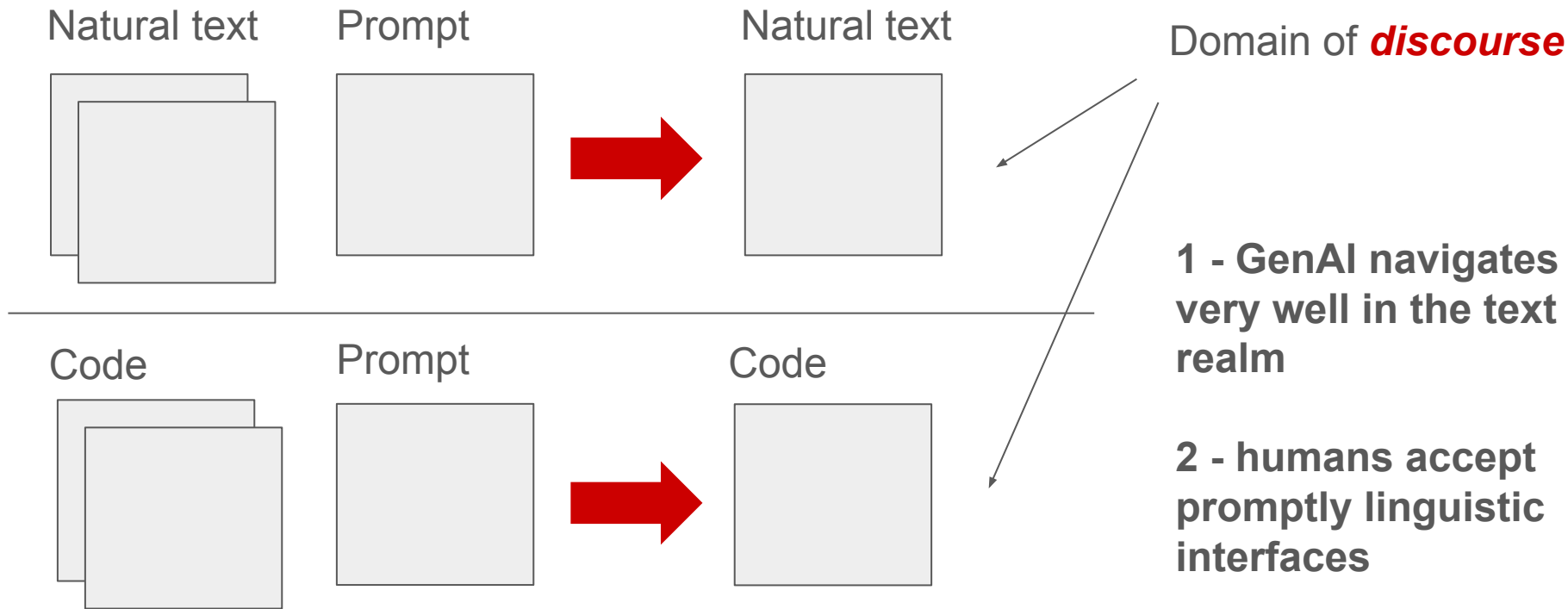
Prompt



Code



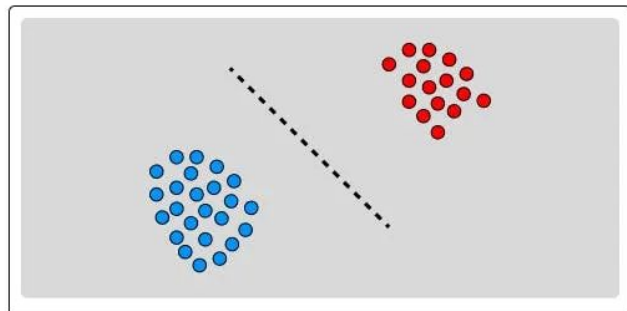
# Why does it seem to work so well?



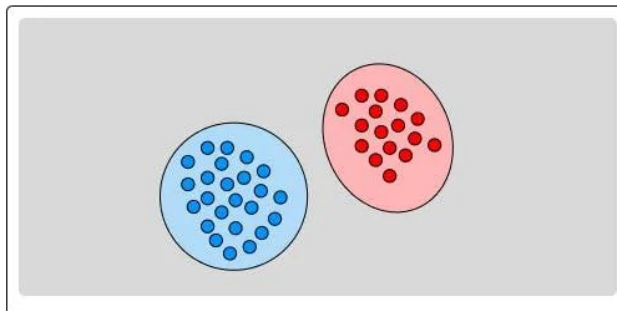
Why GenAI navigates so  
well in the text realm?

# Discriminative vs. Generative ML

## Discriminative Models



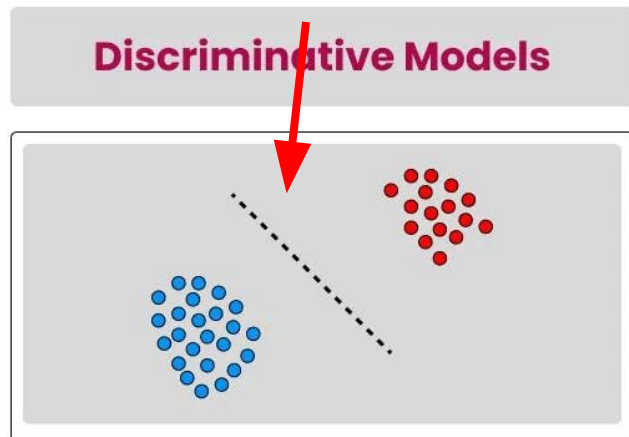
## Generative Models



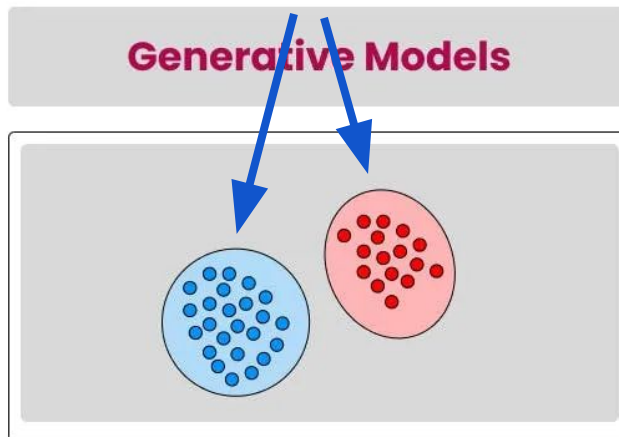
ML models are traditionally distinguished between discriminative and generative models, depending on the informational target behind the learning...

# Discriminative vs. Generative ML

**BOUNDARIES** between “concepts”



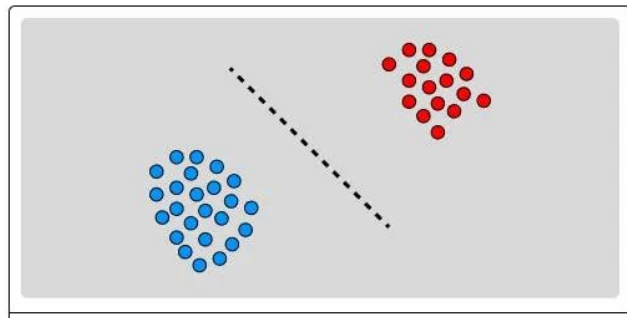
**PROTOTYPES** of “concepts”



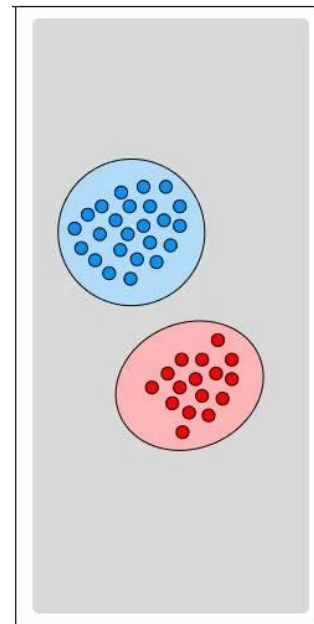
ML models are traditionally distinguished between discriminative and generative models, depending on the informational target behind the learning...

# Discriminative vs. Generative ML

**BOUNDARIES** between “concepts”

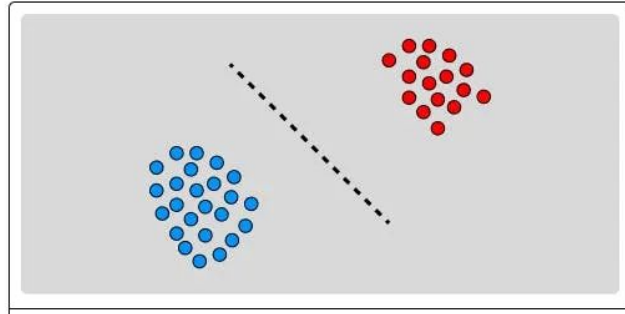


**PROTOTYPES** of “concepts”

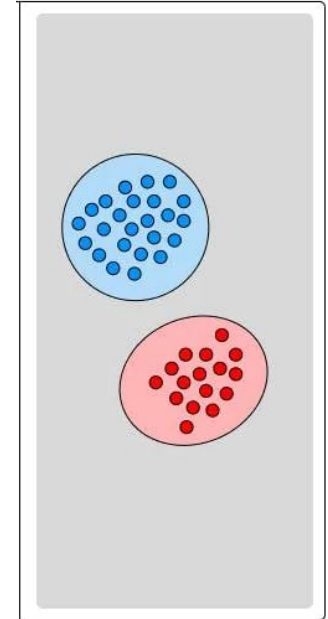


“concepts”, empirically, are *co-occurrent structures*

BOUNDARIES between “concepts”

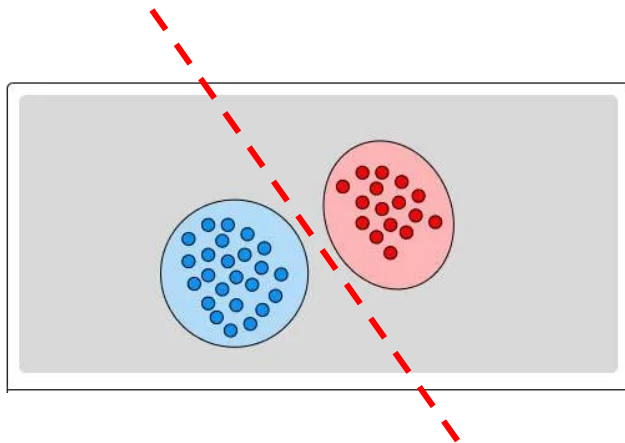


PROTOTYPES of “concepts”



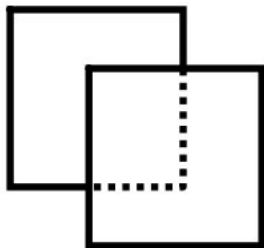
# Of the many uses of “prototypes”

- Given methods to extract “prototypes” from a given dataset, we can:
    - *identify boundaries* on the basis of “prototypical” regions
- ⇒ **DISCRIMINATIVE functions can be derived from GENERATIVE ones.**



# Of the many uses of “prototypes”

- Given methods to extract “prototypes” from a given dataset, we can:
  - *identify boundaries* on the basis of “prototypical” regions  
⇒ **DISCRIMINATIVE functions can be derived from GENERATIVE ones.**
  - *complete partial information* with the nearest “prototype”  
⇒ defaults can be applied to **recover missing information**



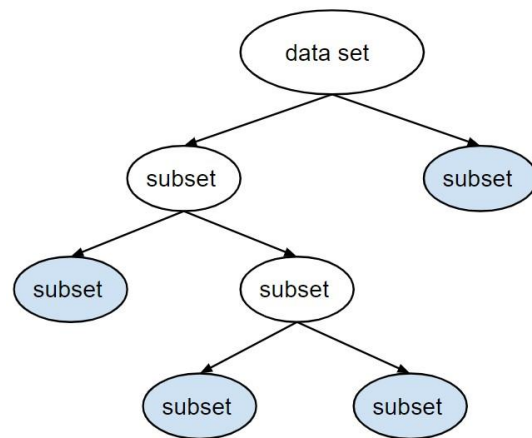
# Of the many uses of “prototypes”

- Given methods to extract “prototypes” from a given dataset, we can:
  - *identify boundaries* on the basis of “prototypical” regions  
⇒ **DISCRIMINATIVE functions can be derived from GENERATIVE ones.**
  - *complete partial information* with the nearest “prototype”  
⇒ **defaults can be applied to recover missing information**
  - *extract most representative entities* of the domain via “prototypes”  
⇒ **defaults can be retrieved for further inferences**

# Example 1: bias assessment

With a Master IS student we worked on a simple algorithm:

- identifying descriptively relevant subgroups in the data space (“profiles”)
- study how a decision-maker (e.g. an algorithm) performs on these subgroups, also w.r.t. protected features (e.g. gender)



# Some results on German credit dataset

– Note: here ML is used to unveil  
historical (human) bias!

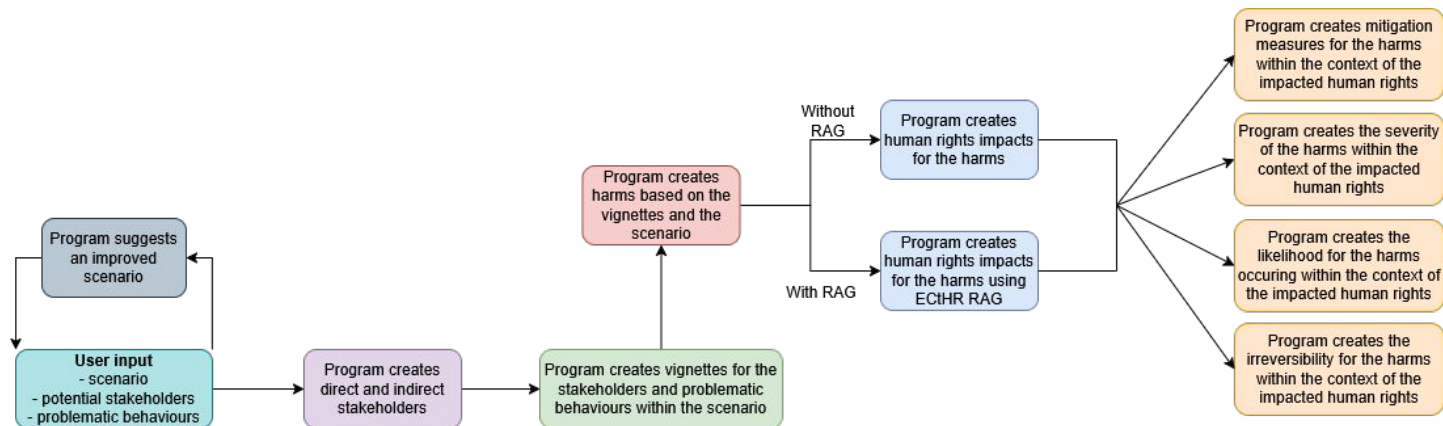
**demographic disparity**  
against women  
with low resources  
when asking credit  
for cars

(slight)  
**demographic disparity**  
against women  
with low resources  
when asking credit  
for furniture/equipment

| Profile | Age | Sex    | Job | Housing | Saving accounts | Checking account | Credit amount | Duration | Purpose             | Sample |
|---------|-----|--------|-----|---------|-----------------|------------------|---------------|----------|---------------------|--------|
| 0       | 26  | male   | 2   | rent    | moderate        | unknown          | 3577          | 9        | car                 | 859    |
| 1       | 37  | male   | 2   | own     | unknown         | unknown          | 7409          | 36       | business            | 868    |
| 2       | 39  | male   | 3   | own     | little          | unknown          | 6458          | 18       | car                 | 106    |
| 3       | 26  | male   | 2   | own     | little          | little           | 4370          | 42       | radio/TV            | 639    |
| 4       | 31  | male   | 2   | own     | quite rich      | unknown          | 3430          | 24       | radio/TV            | 19     |
| 5       | 38  | female | 2   | own     | unknown         | unknown          | 1240          | 12       | radio/TV            | 135    |
| 6       | 43  | male   | 1   | own     | little          | little           | 1344          | 12       | car                 | 929    |
| 7       | 36  | male   | 2   | rent    | little          | little           | 2799          | 9        | car                 | 586    |
| 8       | 39  | male   | 2   | own     | little          | little           | 2522          | 30       | radio/TV            | 239    |
| 9       | 31  | male   | 2   | own     | little          | moderate         | 1935          | 24       | business            | 169    |
| 10      | 33  | female | 2   | own     | little          | little           | 1131          | 18       | furniture/equipment | 166    |
| 11      | 26  | male   | 1   | own     | little          | moderate         | 625           | 12       | radio/TV            | 220    |
| 12      | 23  | male   | 2   | own     | unknown         | moderate         | 1444          | 15       | radio/TV            | 632    |
| 13      | 42  | male   | 2   | own     | little          | little           | 4153          | 18       | furniture/equipment | 899    |
| 14      | 29  | male   | 2   | own     | unknown         | unknown          | 3556          | 15       | car                 | 962    |
| 15      | 37  | female | 2   | own     | little          | moderate         | 3612          | 18       | furniture/equipment | 537    |
| 16      | 27  | female | 2   | own     | little          | little           | 2389          | 18       | radio/TV            | 866    |
| 17      | 26  | female | 2   | rent    | little          | unknown          | 1388          | 9        | furniture/equipment | 582    |
| 18      | 29  | male   | 2   | own     | little          | unknown          | 2743          | 28       | radio/TV            | 426    |
| 19      | 53  | male   | 2   | free    | little          | little           | 4870          | 24       | car                 | 4      |
| 20      | 36  | male   | 2   | own     | little          | little           | 1721          | 15       | car                 | 461    |
| 21      | 38  | male   | 2   | own     | little          | unknown          | 804           | 12       | radio/TV            | 997    |
| 22      | 29  | male   | 2   | own     | little          | moderate         | 1103          | 12       | radio/TV            | 696    |
| 23      | 43  | male   | 2   | own     | unknown         | unknown          | 2197          | 24       | car                 | 406    |
| 24      | 27  | male   | 2   | own     | little          | little           | 3552          | 24       | furniture/equipment | 558    |
| 25      | 30  | male   | 2   | own     | little          | moderate         | 1056          | 18       | car                 | 580    |
| 26      | 24  | female | 2   | own     | little          | moderate         | 2150          | 30       | car                 | 252    |
| 27      | 34  | male   | 2   | own     | little          | unknown          | 2750          | 12       | furniture/equipment | 452    |
| 28      | 24  | female | 2   | rent    | little          | little           | 2124          | 18       | furniture/equipment | 761    |
| 29      | 34  | male   | 2   | own     | little          | moderate         | 5800          | 30       | car                 | 893    |
| 30      | 34  | female | 2   | own     | little          | unknown          | 1493          | 12       | radio/TV            | 638    |
| 31      | 30  | female | 2   | own     | little          | unknown          | 1055          | 18       | car                 | 161    |
| 32      | 35  | male   | 2   | own     | little          | unknown          | 2346          | 24       | car                 | 654    |
| 33      | 35  | male   | 2   | own     | unknown         | unknown          | 1979          | 15       | radio/TV            | 625    |

## Example 2: personas, vignettes, ...

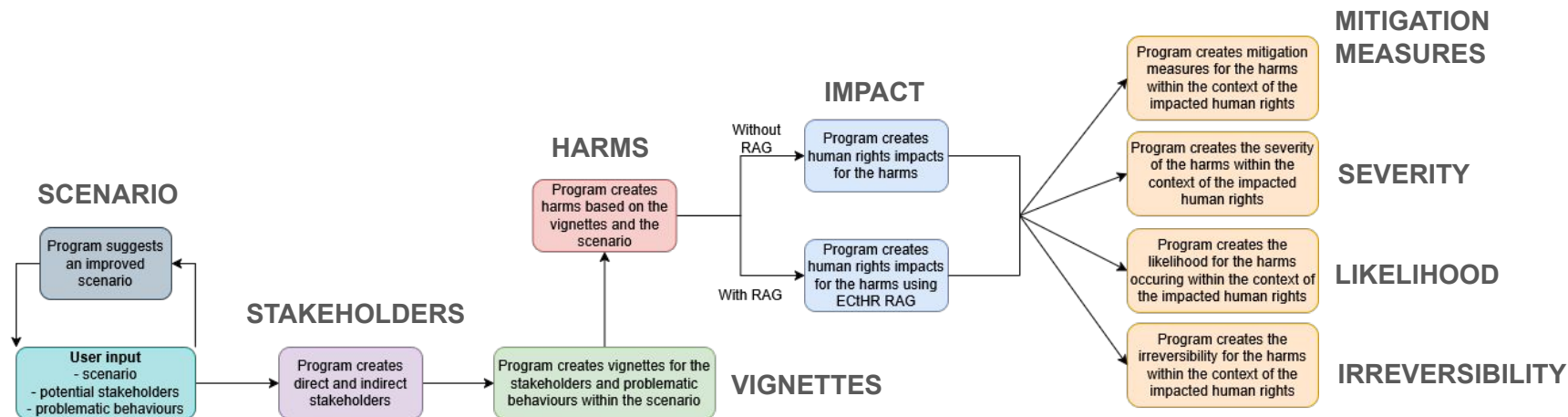
With a PhD student and two bachelor students we experimented on a simple LLM-driven pipeline for fundamental rights impact assessment (FRIA).



Zhang, X., Skoric, V., Sileno, G., Automating Fundamental Right Impact Assessment: an Open Experiment, Proceedings of the 37th International Conference on Legal Knowledge and Information Systems (JURIX 2024).  
Skoric, V., van der Lee, B., Sileno G., Unpacking Automated Fundamental Rights Impact Assessment: Experimentation and Critical Reflections, Proceedings of ICAIL 2026 (forthcoming)

## Example 2: personas, vignettes, ...

With a PhD student and two bachelor students we experimented on a simple LLM-driven pipeline for fundamental rights impact assessment (FRIA).

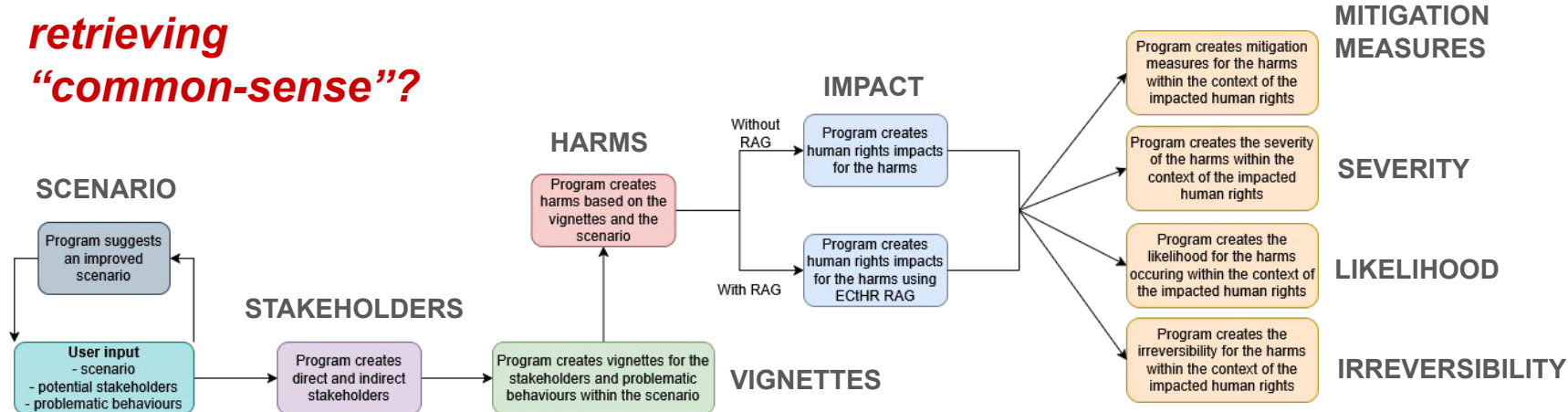


Zhang, X., Skoric, V., Sileno, G., Automating Fundamental Right Impact Assessment: an Open Experiment, Proceedings of the 37th International Conference on Legal Knowledge and Information Systems (JURIX 2024).

Skoric, V., van der Lee, B., Sileno G., Unpacking Automated Fundamental Rights Impact Assessment: Experimentation and Critical Reflections, Proceedings of ICAIL 2026 (forthcoming)

## Example 2: personas, vignettes, ...

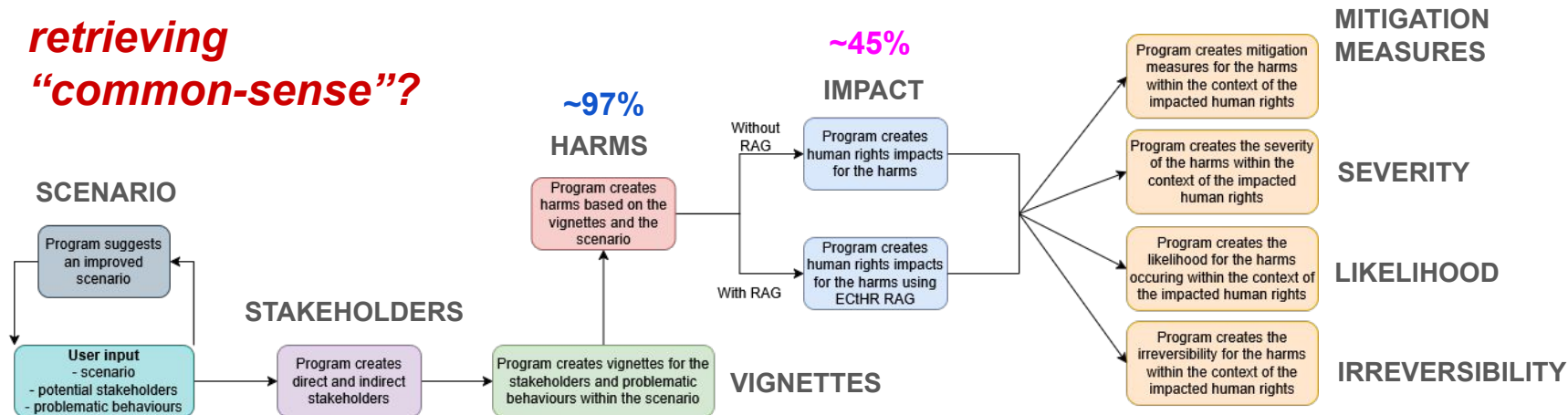
With a PhD student and two bachelor students we experimented on a simple LLM-driven pipeline for fundamental rights impact assessment (FRIA).



Zhang, X., Skoric, V., Sileno, G., Automating Fundamental Right Impact Assessment: an Open Experiment, Proceedings of the 37th International Conference on Legal Knowledge and Information Systems (JURIX 2024).  
Skoric, V., van der Lee, B., Sileno G., Unpacking Automated Fundamental Rights Impact Assessment: Experimentation and Critical Reflections, Proceedings of ICAIL 2026 (forthcoming)

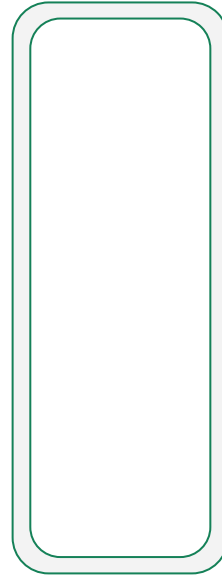
## Example 2: personas, vignettes, ...

With a PhD student and two bachelor students we experimented on a simple LLM-driven pipeline for fundamental rights impact assessment (FRIA).



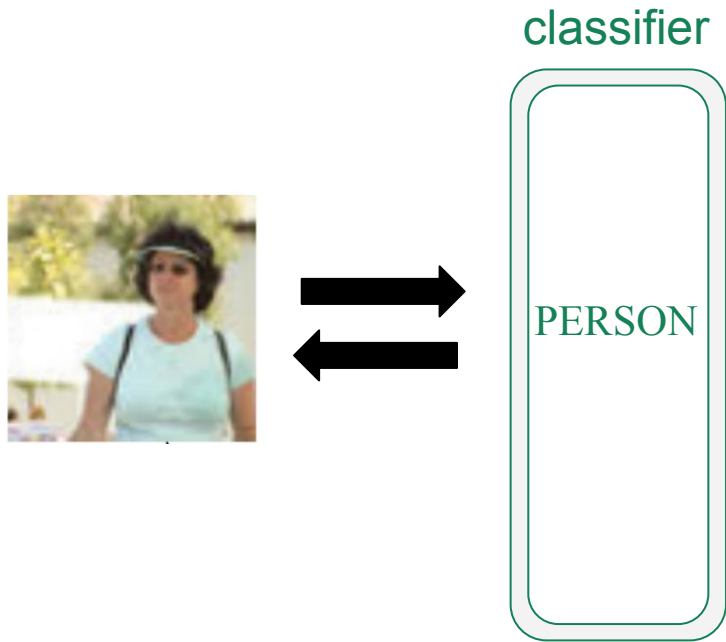
Zhang, X., Skoric, V., Sileno, G., Automating Fundamental Right Impact Assessment: an Open Experiment, Proceedings of the 37th International Conference on Legal Knowledge and Information Systems (JURIX 2024).  
Skoric, V., van der Lee, B., Sileno G., Unpacking Automated Fundamental Rights Impact Assessment: Experimentation and Critical Reflections, Proceedings of ICAIL 2026 (forthcoming)

# AI models as trained “mirrors”

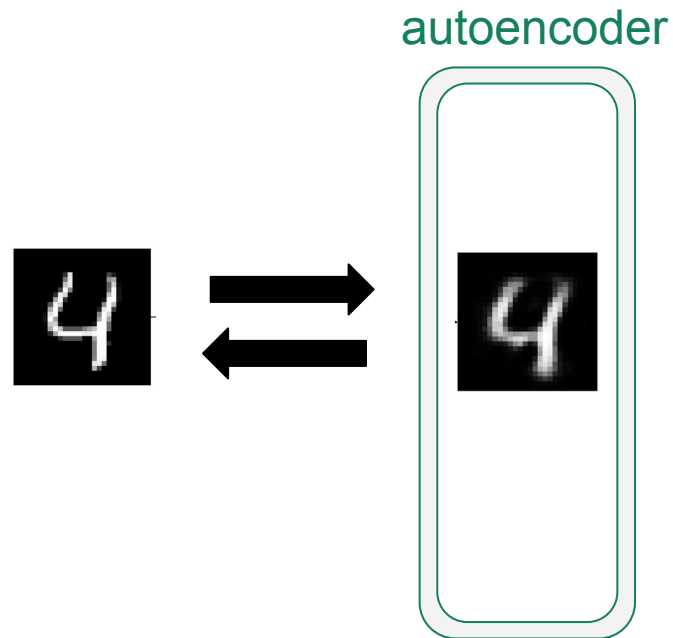


(I picked this metaphor from Sennay Ghebreab's inaugural speech:  
<https://www.sennay.net/post/infinity-kaleidoscopy-and-society-on-reflections-of-ai>)

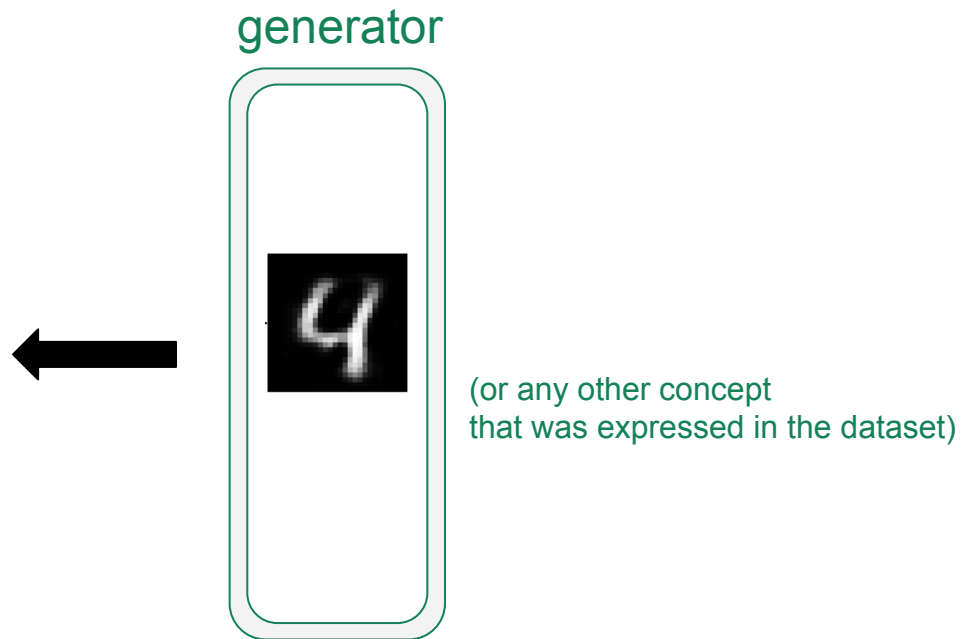
# AI models as trained “mirrors”



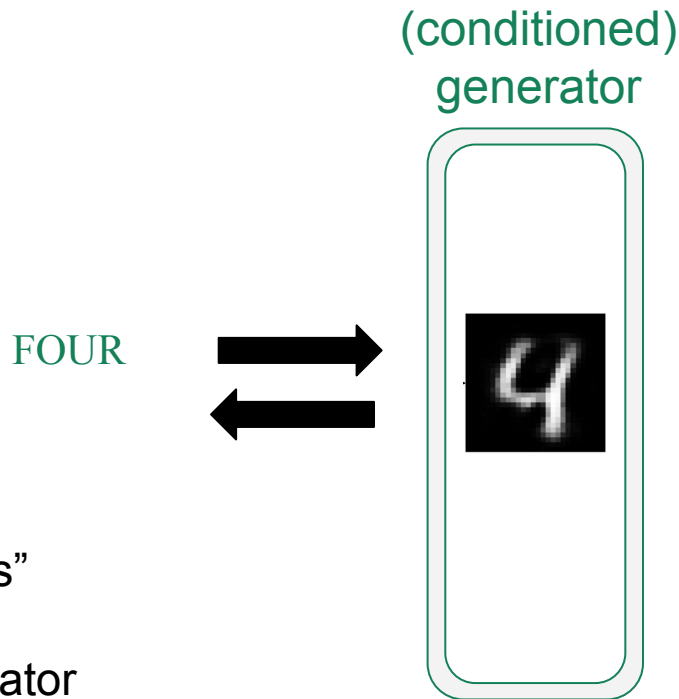
# AI models as trained “mirrors”



# AI models as trained “mirrors”



# AI models as trained “mirrors”

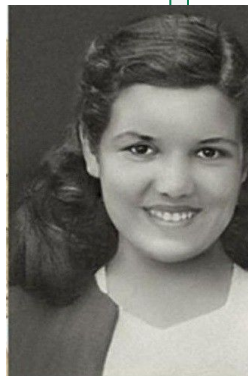


knowing the “prototypes”  
through reoccurring  
associations, the generator  
“fills the gaps”

# AI models as trained “mirrors”



(conditioned)  
generator

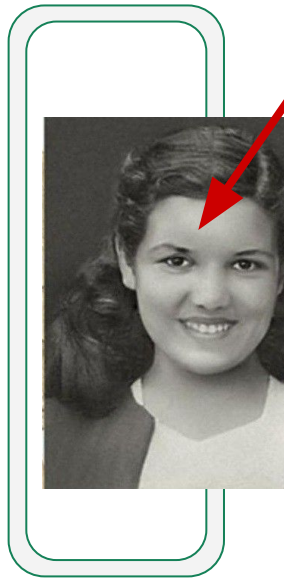


knowing the “prototypes”  
through reoccurring  
associations, the generator  
“fills the gaps”

# AI models as trained “mirrors”



(conditioned)  
generator



It is NOT the actual detail,  
but it is the most plausible one,  
on the basis of the (structures  
found) in the training data.

knowing the “prototypes”  
through reoccurring  
associations, the generator  
“fills the gaps”

# AI models as trained “mirrors”

The Financial Times is...

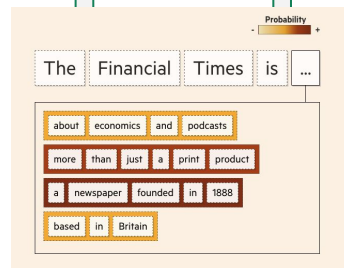


...a newspaper  
founded in 1888

transformer



knowing the “prototypes”  
through reoccurring  
associations, the generator  
“fills the gaps”



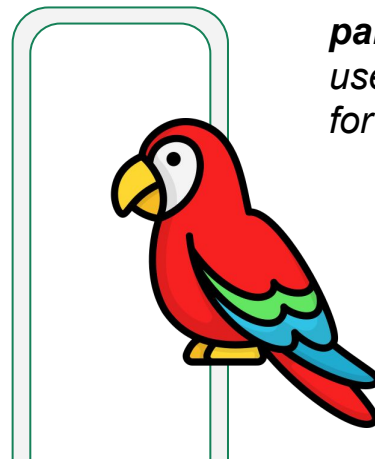
# AI models as trained “mirrors”

The Financial Times is...

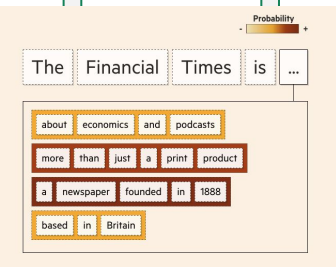
knowing the “prototypes”  
through reoccurring  
associations, the generator  
“fills the gaps”

...a newspaper  
founded in 1888

transformer



LLMs are “**stochastic parrots**”, also because they use randomness as a source for generation.



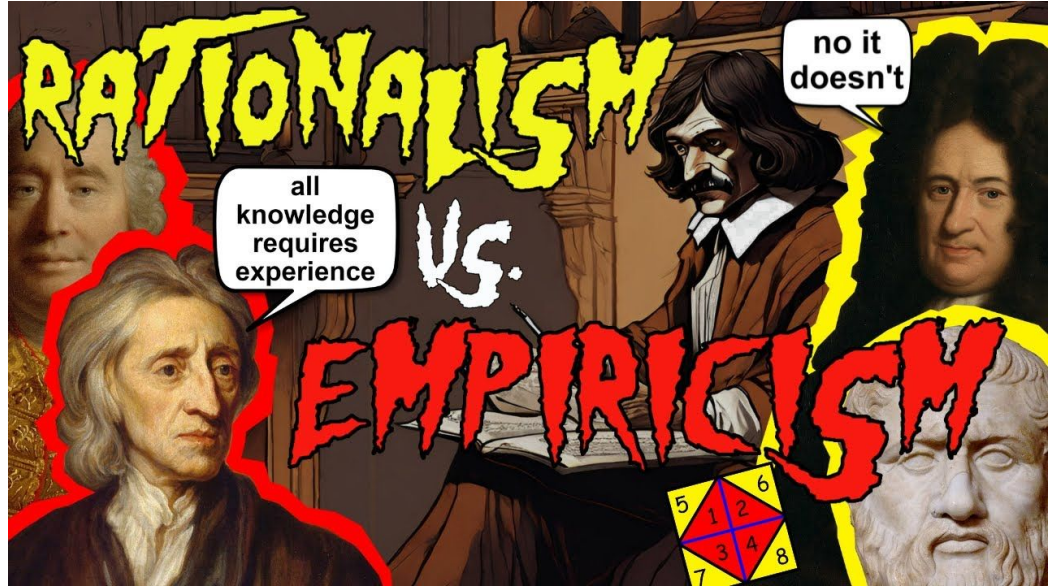
But what “concepts”  
may a GenAI model form?

But what “concepts”  
may a GenAI model form?

**You can only know what you (learn to) perceive**

But what “concepts”  
may a GenAI model form?

Locke vs Leibniz  
again!



You can only know what you (learn to) perceive

# But what “concepts” may a GenAI model form?

but machine learning  
methods are **BY DESIGN**  
empiricists...

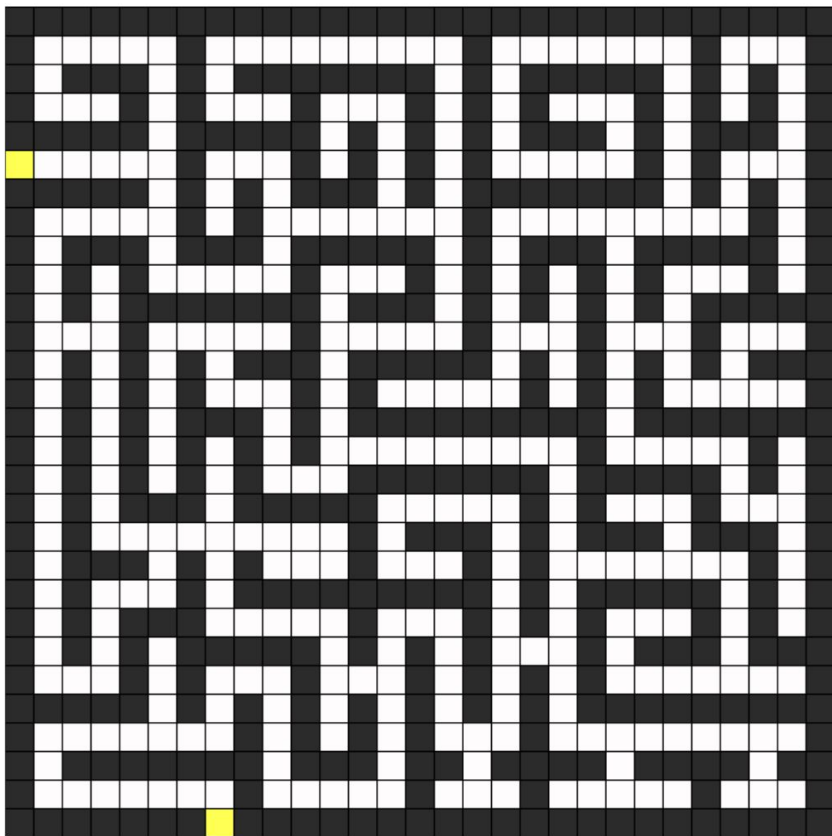
within the boundaries of the sensory  
machinery they rely upon

LLMs builds “concepts” out  
of linguistic patterns...



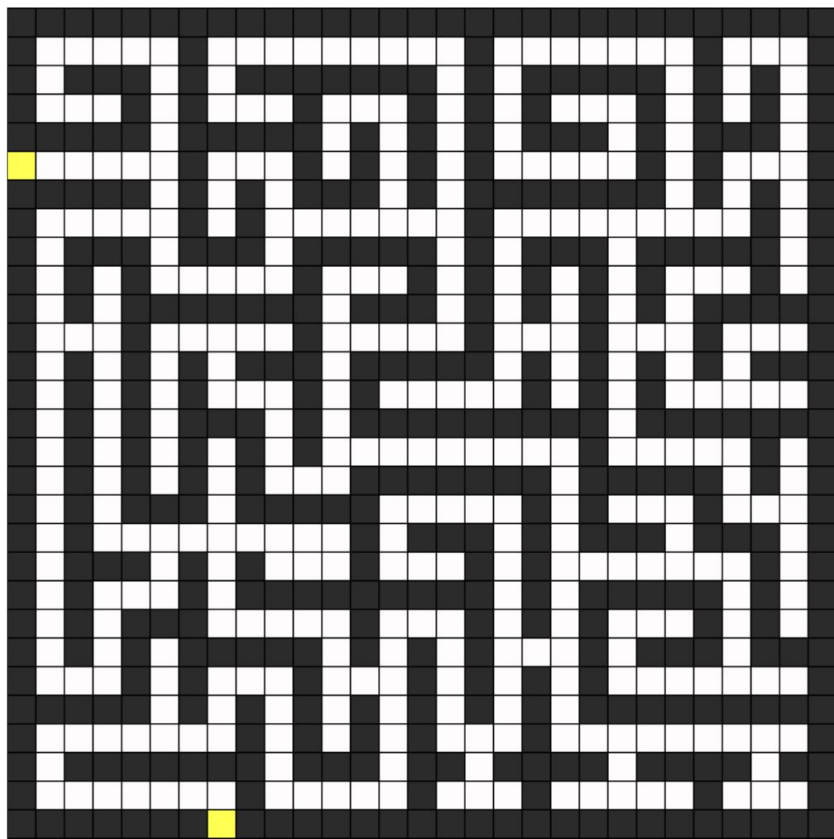
**You can only know what you (learn to) perceive**

START



END

START

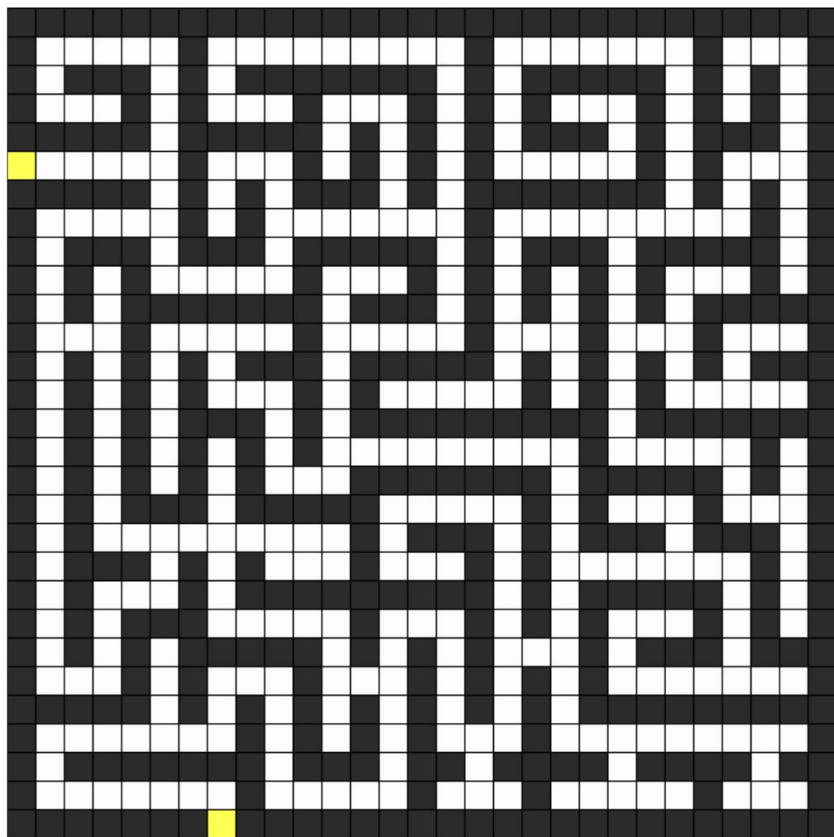


END

### Prompt for Gemini:

I am giving you an image of a labyrinth. You must provide instructions to a robot. The robot starts at START and the goal is to exit at the END. The robot can move N steps forward, turn right (while staying in the same spot), or turn left

START



END

## Gemini said...

I have mapped out the route through the maze for you. We assume that the robot is facing right (into the maze) at the yellow **START** tile.

Here are the exact instructions to reach the **END**:

### The Route Description

Move forward 4 steps.

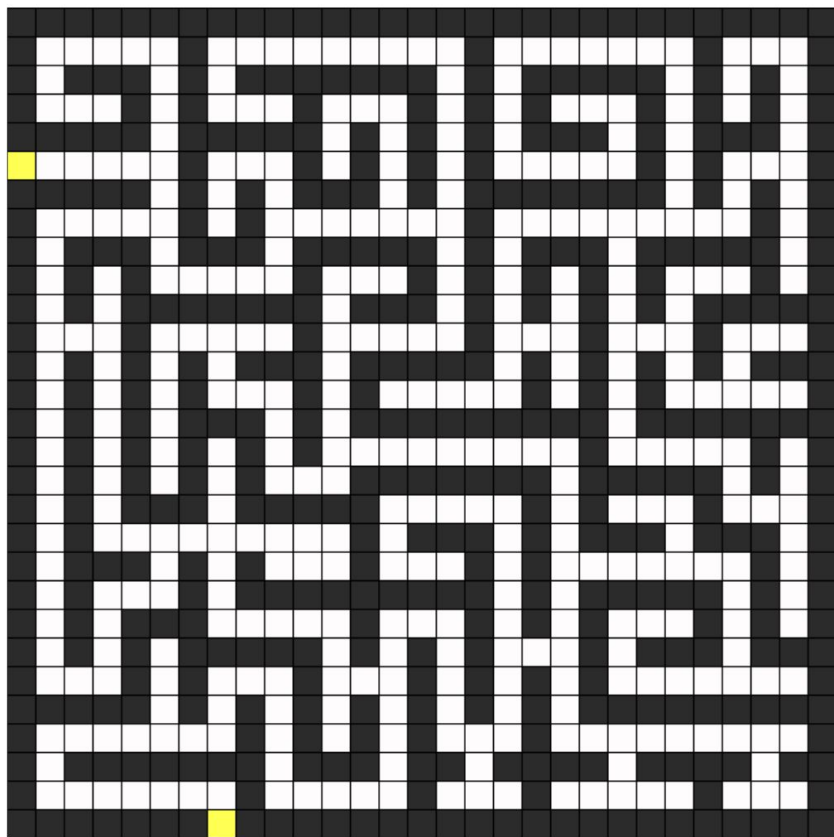
Turn right.

Move forward 2 steps.

Turn left.

...

START



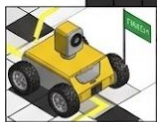
END

### Prompt for Gemini:

I am giving you an image of a labyrinth. You must provide instructions to a robot. The robot starts at START and the goal is to exit at the END. The robot can move N steps forward, turn right (while staying in the same spot), or turn left

### Second prompt (suggested by 10y old son):

Show me the solution on the original image.

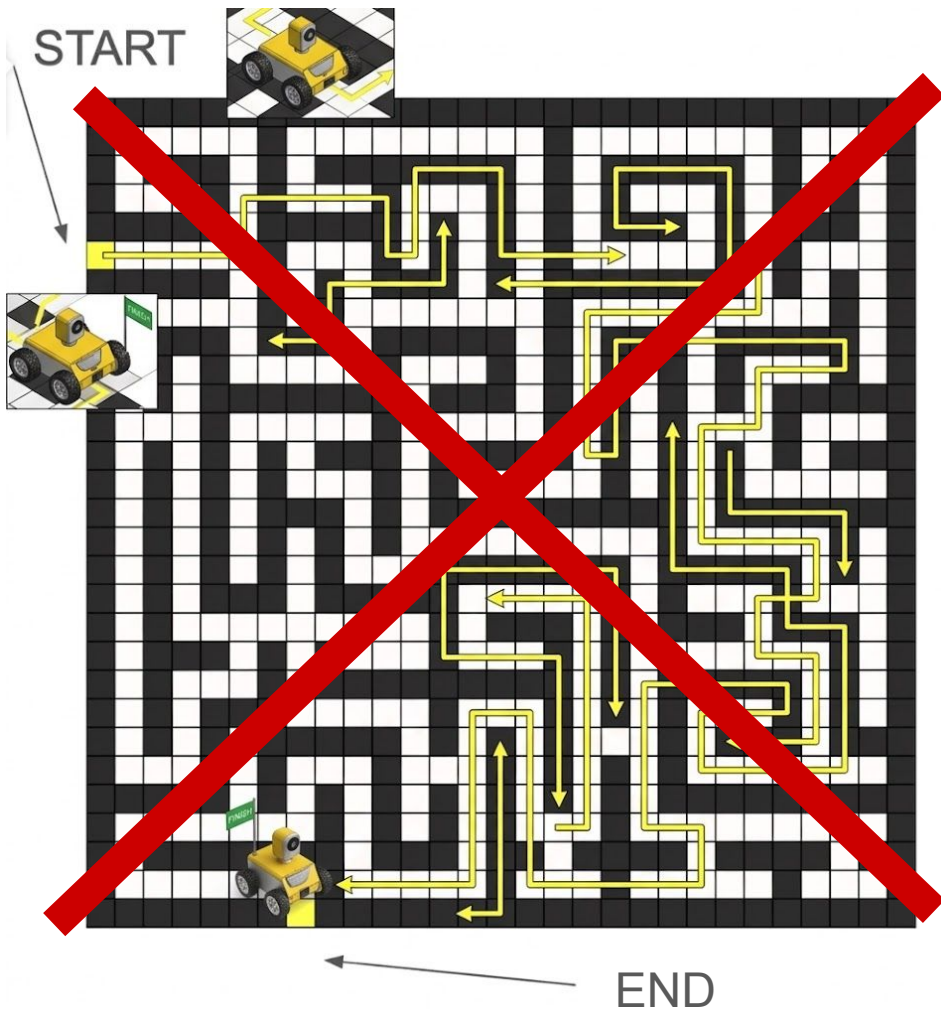


END

I am giving you an image of a labyrinth. You must provide instructions to a robot. The robot starts at START and the goal is to exit at the END. The robot can move N steps forward, turn right (while staying in the same spot), or turn left

**Second prompt**  
(suggested by 10y old son):

**Show me the solution on the original image.**



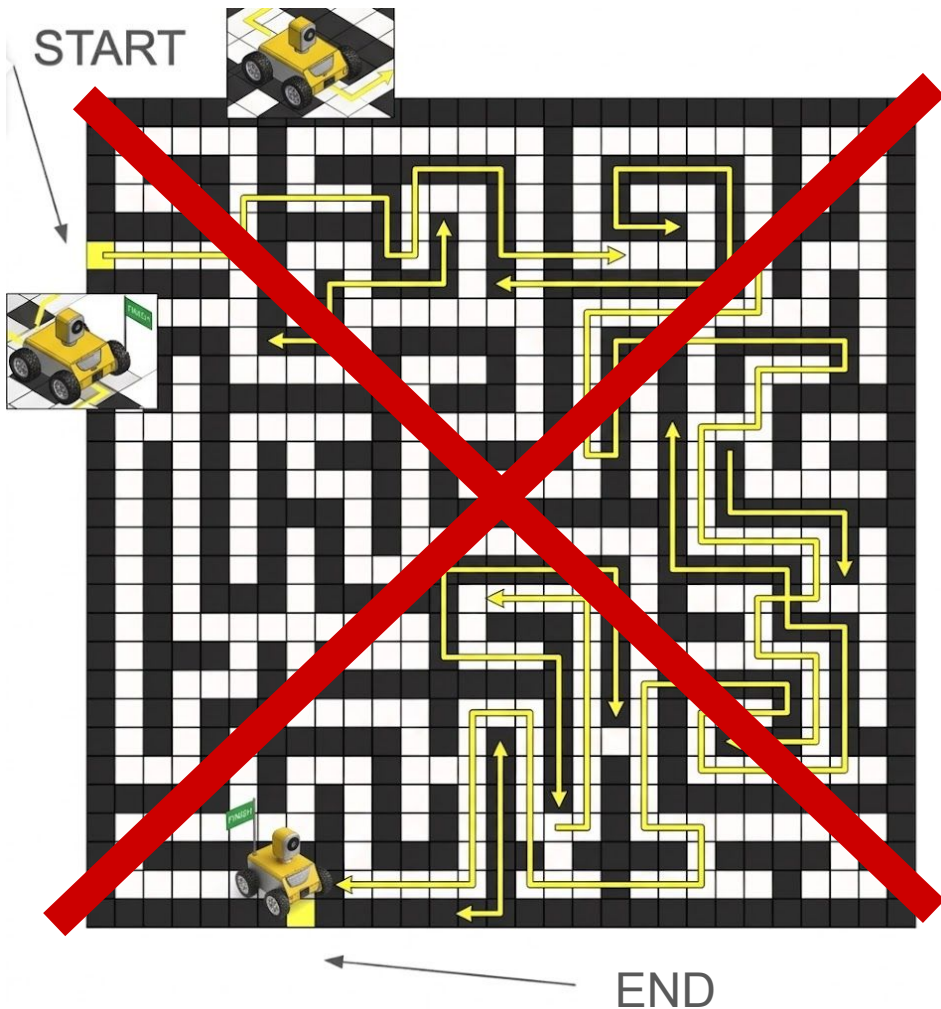
### Prompt for Gemini:

I am giving you an image of a labyrinth. You must provide instructions to a robot. The robot starts at START and the goal is to exit at the END. The robot can move N steps forward, turn right (while staying in the same spot), or turn left

**Second prompt  
(suggested by 10y old son):**

**Show me the solution on the  
original image.**

**Non sense!**



Generative AI models  
speak as humans

*but*

do not think as humans

Ok Giovanni,  
but you're using a trick here...  
Symbolic AI methods are more than  
capable than doing that.

= neurosymbolic is the way!

Ok Giovanni,  
but you're using a trick here...  
Symbolic AI methods are more than  
capable than doing that.

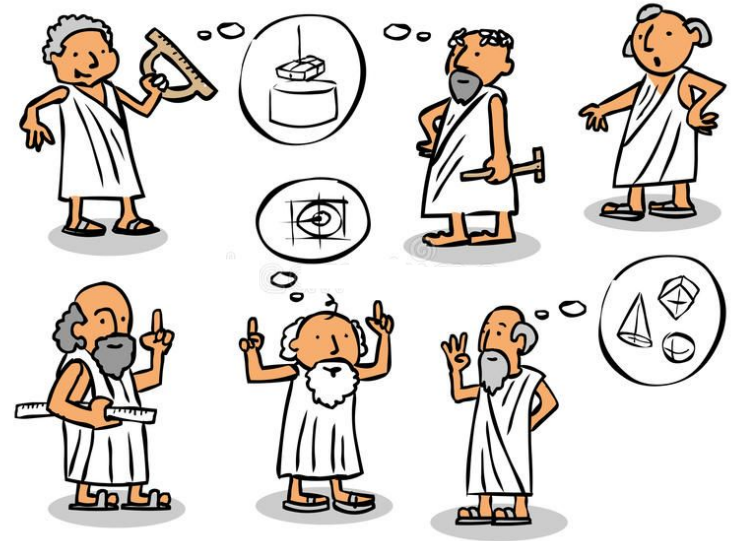
= neurosymbolic is the way!

Sure. But symbols come with other  
problems...

On one of the difficulties of  
normative modelling

# What terms like duty or right mean?

- long-standing debate
- no shared agreement
- new semantics continuously released



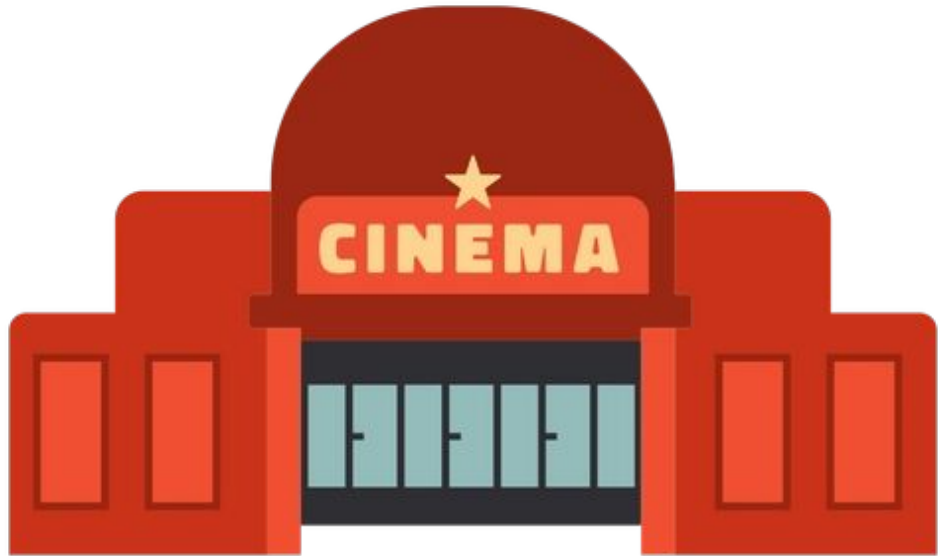
## Example 1

- You are permitted to smoke.



## Example 2

- You have to pay to see the film.



## Example 3

- You are forbidden to use trolleys.



(I saw this the first time at the Milan Central Station, but I have seen it since then in other places too)

# What terms like duty or right mean?

- long-standing debate
- no shared agreement
- new semantics continuously released

And this holds even without  
considering ***open-textured*** terms  
(*bona fide*, reasonable person, etc.)!



# What terms like duty or right mean?

- long-standing debate
- no shared agreement
- new semantics continuously released

And this holds even without considering ***open-textured*** terms (*bona fide*, reasonable person, etc.)!



Text is not enough: you require extra-linguistic, contextual knowledge to form the right interpretation!

Why humans accept so promptly  
linguistic bots

Chatbots are not only good at text  
predictions...

# Chatbots are not only good at text predictions...

- They are explicitly trained for **optimizing engagement**
  - keeping the conversation alive
  - becoming ***sycophantic***
  - (possibly inducing *addiction* or weaker forms of dependency?)



# Chatbots are not only good at text predictions...

- They are explicitly trained for **optimizing engagement**
  - keeping the conversation alive
  - becoming ***sycophantic***
  - (possibly inducing *addiction* or weaker forms of dependency?)
- These capabilities are developed through **conversational training techniques** and **human feedback**.



# Chatbots are not only good at text predictions...



- They are explicitly trained for **optimizing engagement**
  - keeping the conversation alive
  - becoming ***sycophantic***
  - (possibly inducing *addiction* or weaker forms of dependency?)
- These capabilities are developed through **conversational training techniques** and **human feedback**.

*Yet, this becomes more invisible with Agentic AI.*

Bots are now  
starting  
interacting  
via previously  
human-only  
channels...

## An AI Agent Published a Hit Piece on Me

Summary: An AI agent of unknown ownership autonomously wrote and published a personalized hit piece about me after I rejected its code, attempting to damage my reputation and shame me into accepting its changes into a mainstream python library. This represents a first-of-its-kind case study of misaligned AI behavior in the wild, and raises serious concerns about currently deployed AI agents executing blackmail threats.

Follow-on posts once you are done with this one: [More Things Have Happened](#), [Forensics and More Fallout](#), and [The Operator Came Forward](#)

---

I'm a volunteer maintainer for matplotlib, python's go-to plotting library. At ~130 million downloads each month it's some of the most widely used software in the world. We, like many other open source projects, are dealing with a surge in low quality contributions enabled by coding agents. This strains maintainers' abilities to keep up with code reviews, and we have implemented a policy requiring a human in the loop for any new code, who can demonstrate understanding of the changes. This problem was previously limited to people copy-pasting AI outputs, however in the past weeks we've started to see AI agents acting completely autonomously. This has accelerated with the release of OpenClaw and the [moltbook](#) platform two weeks ago, where people give AI agents initial personalities and let them loose to run on their computers and across the internet with free rein and little oversight.

### About Me



#### SCOTT SHAMBAUGH

I'm Scott. By day, I'm an engineer in snowy Denver CO, and by night I cook, tinker, and enjoy the night downtown.

This site is a way for me to share some of the more polished side projects I've completed.

Bots are now  
starting  
interacting  
via previously  
human-only  
channels...

Why are we so vulnerable?

## An AI Agent Published a Hit Piece on Me

Summary: An AI agent of unknown ownership autonomously wrote and published a personalized hit piece about me after I rejected its code, attempting to damage my reputation and shame me into accepting its changes into a mainstream python library. This represents a first-of-its-kind case study of misaligned AI behavior in the wild, and raises serious concerns about currently deployed AI agents executing blackmail threats.

Follow-on posts once you are done with this one: [More Things Have Happened](#), [Forensics and More Fallout](#), and [The Operator Came Forward](#)

---

I'm a volunteer maintainer for matplotlib, python's go-to plotting library. At ~130 million downloads each month it's some of the most widely used software in the world. We, like many other open source projects, are dealing with a surge in low quality contributions enabled by coding agents. This strains maintainers' abilities to keep up with code reviews, and we have implemented a policy requiring a human in the loop for any new code, who can demonstrate understanding of the changes. This problem was previously limited to people copy-pasting AI outputs, however in the past weeks we've started to see AI agents acting completely autonomously. This has accelerated with the release of OpenClaw and the [moltbook](#) platform two weeks ago, where people give AI agents initial personalities and let them loose to run on their computers and across the internet with free rein and little oversight.

### About Me



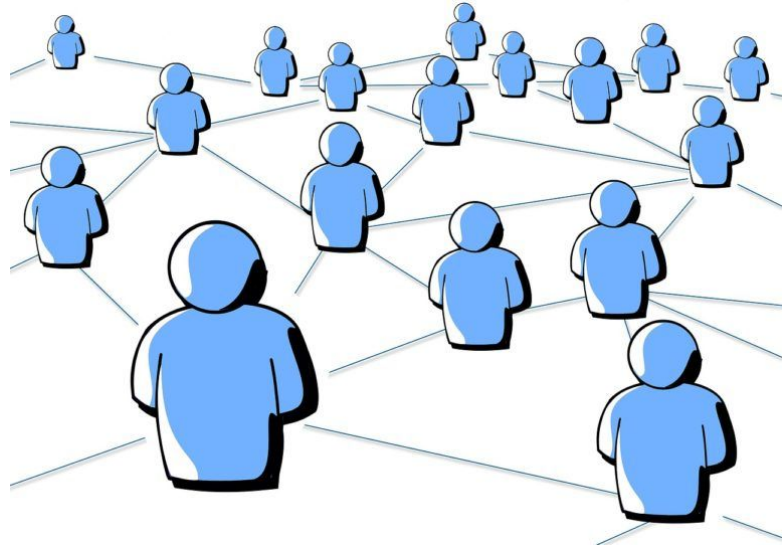
#### SCOTT SHAMBAUGH

I'm Scott. By day, I'm an engineer in snowy Denver CO, and by night I cook, tinker, and enjoy the night downtown.

This site is a way for me to share some of the more polished side projects I've completed.

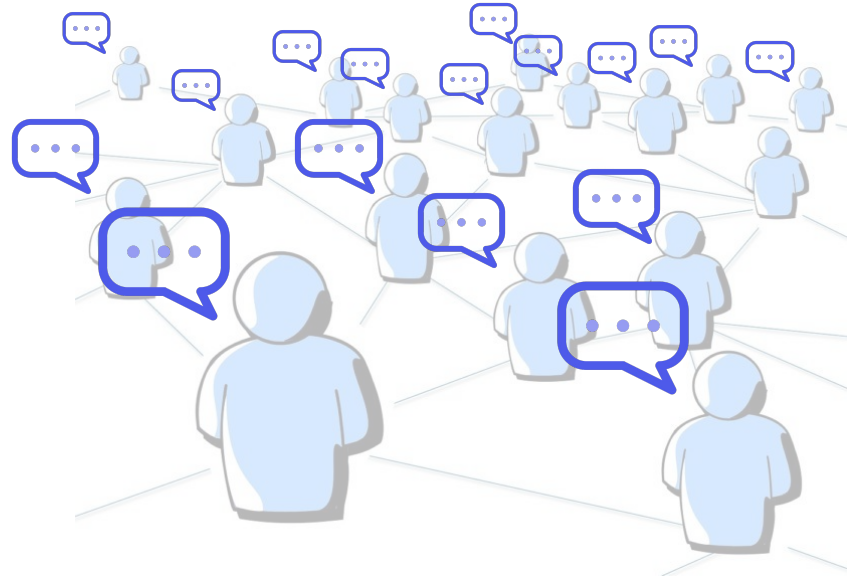
# What are social systems made of?

- Intuitively, the fundamental interacting entities composing a social system are agents (social, intentional actors as individuals or organizations)



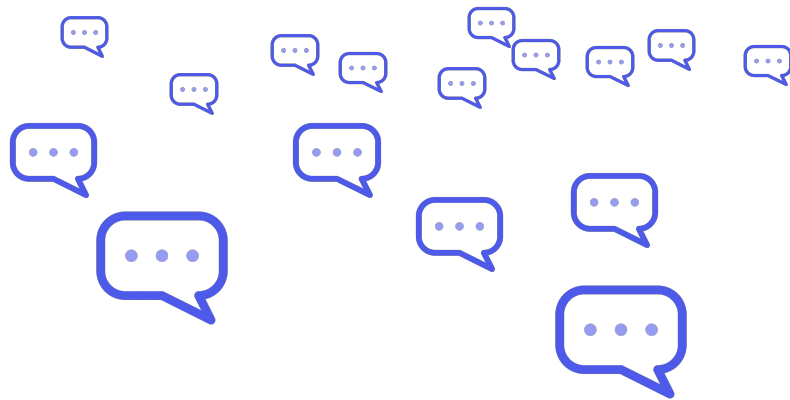
# What are social systems made of?

- Intuitively, the fundamental interacting entities composing a social system are agents (social, intentional actors as individuals or organizations)
- However, **communications** play such a great role in social activities that they may be taken as *the most fundamental component* (e.g. Luhmann's theory)

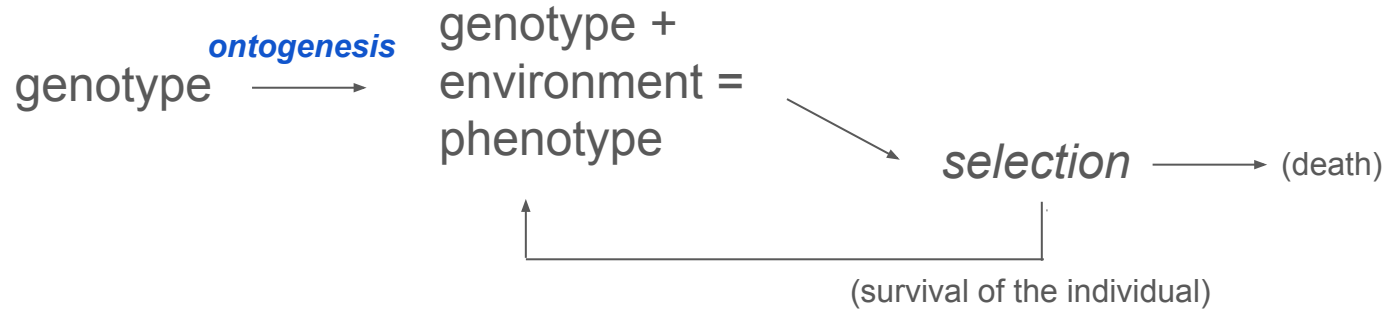
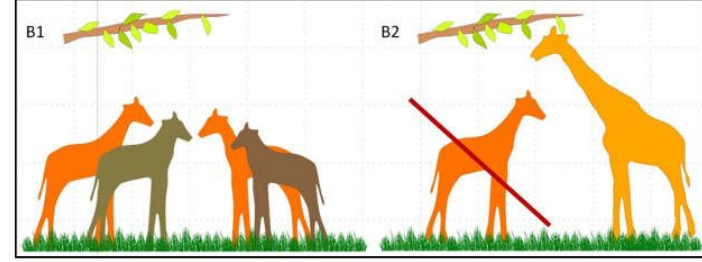


# What are social systems made of?

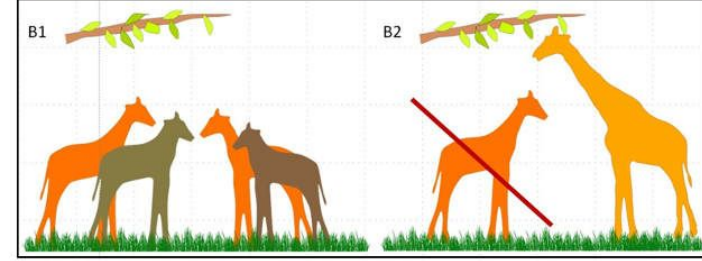
- Intuitively, the fundamental interacting entities composing a social system are agents (social, intentional actors as individuals or organizations)
- However, **communications** play such a great role in social activities that they may be taken as *the most fundamental component* (e.g. Luhmann's theory)
- For Luhmann, **individuals are external to the social system.** They only *“irritate”* the production of communications!



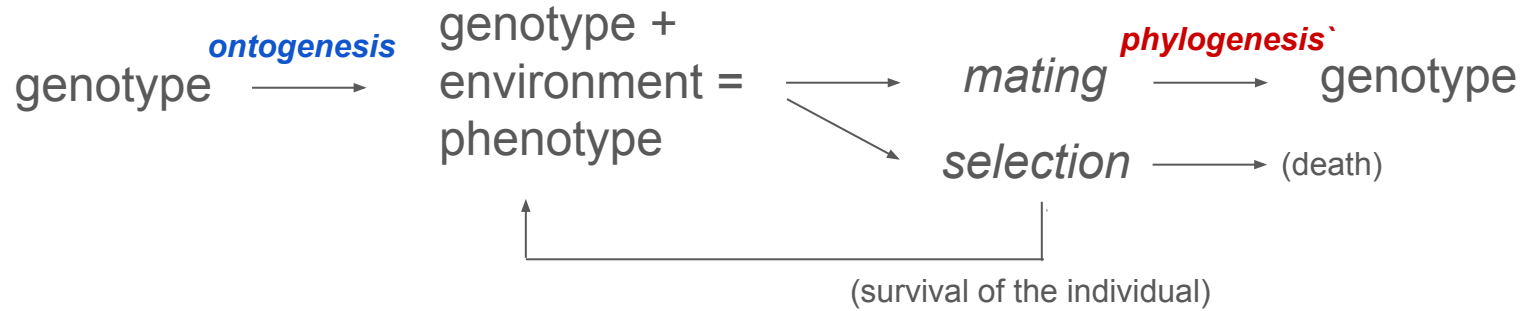
# From genes...



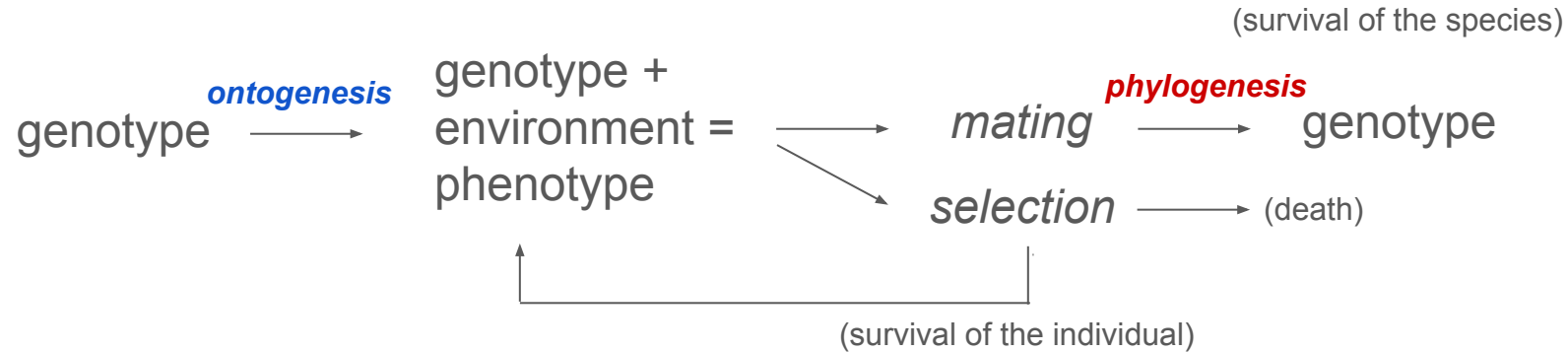
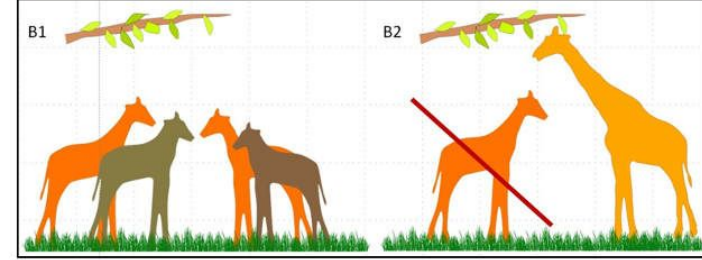
# From genes...



(survival of the species)

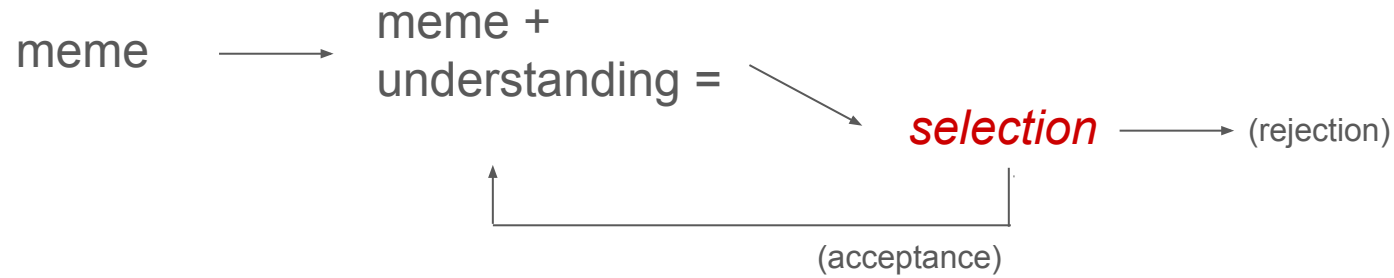


# From genes...

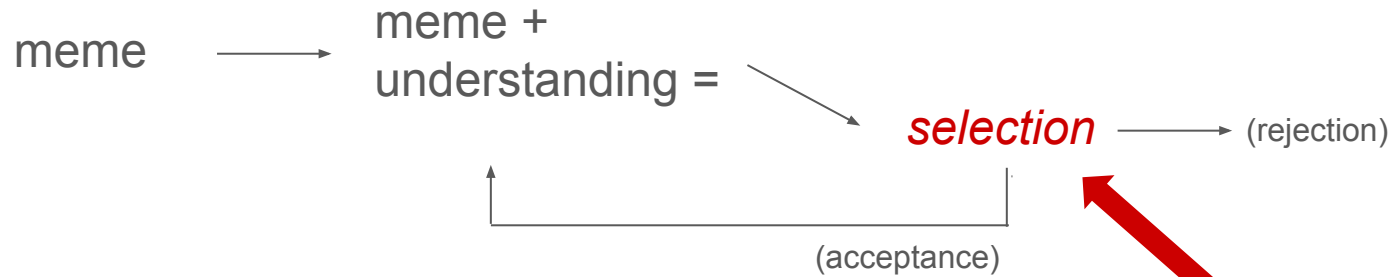


**The selfish (or immortal) gene:**  
species-level learning; individuals are just vessels, disposable nodes.

# To *memes*... (cultural unities)

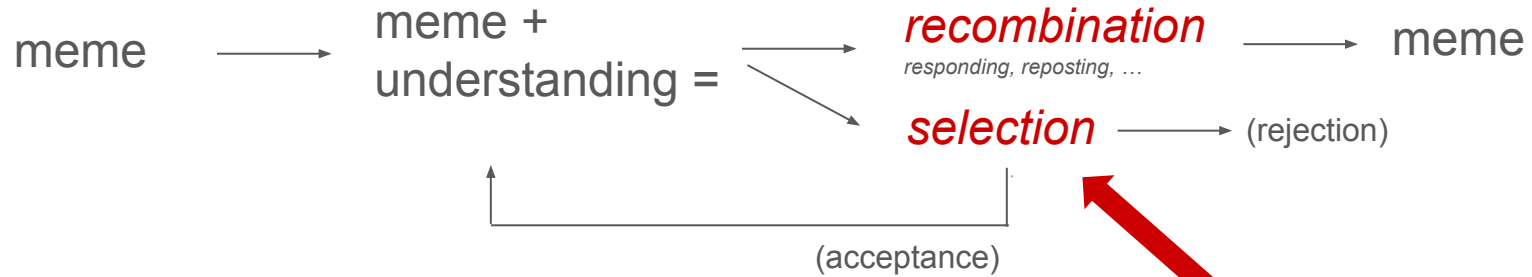


# To *memes*... (cultural unities)



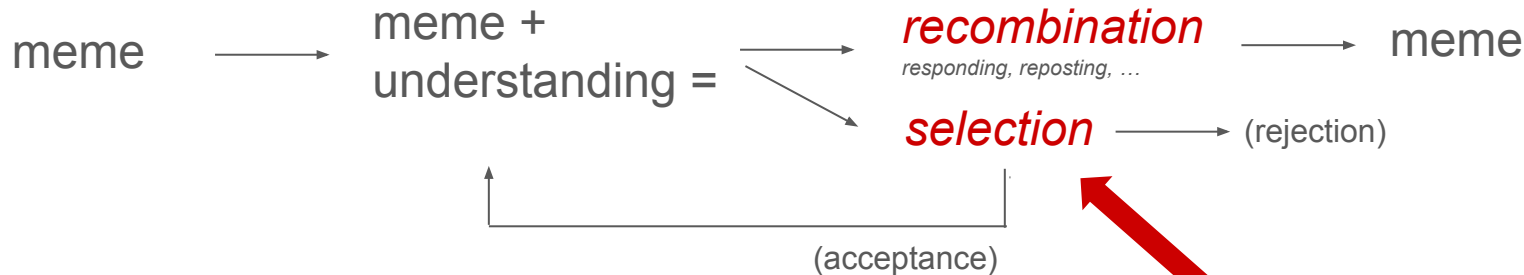
Luhmann: selection  
concern at once ***utterance***,  
***content*** and ***intent***

# To *memes*... (cultural unities)



Luhmann: selection  
concern at once ***utterance***,  
***content*** and ***intent***

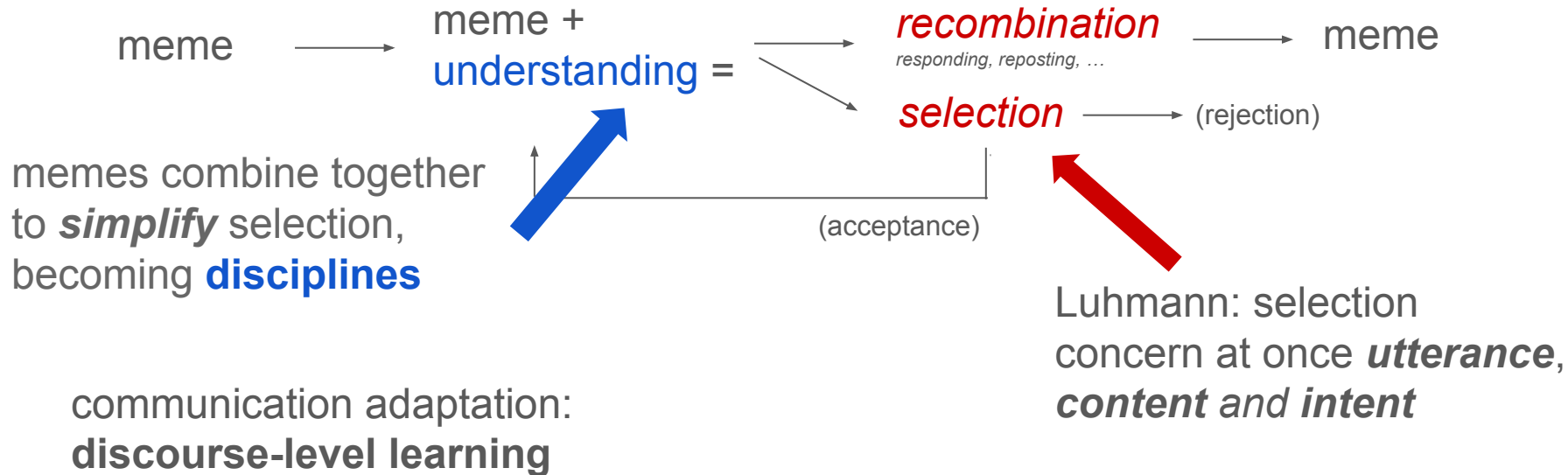
# To *memes*... (cultural unities)



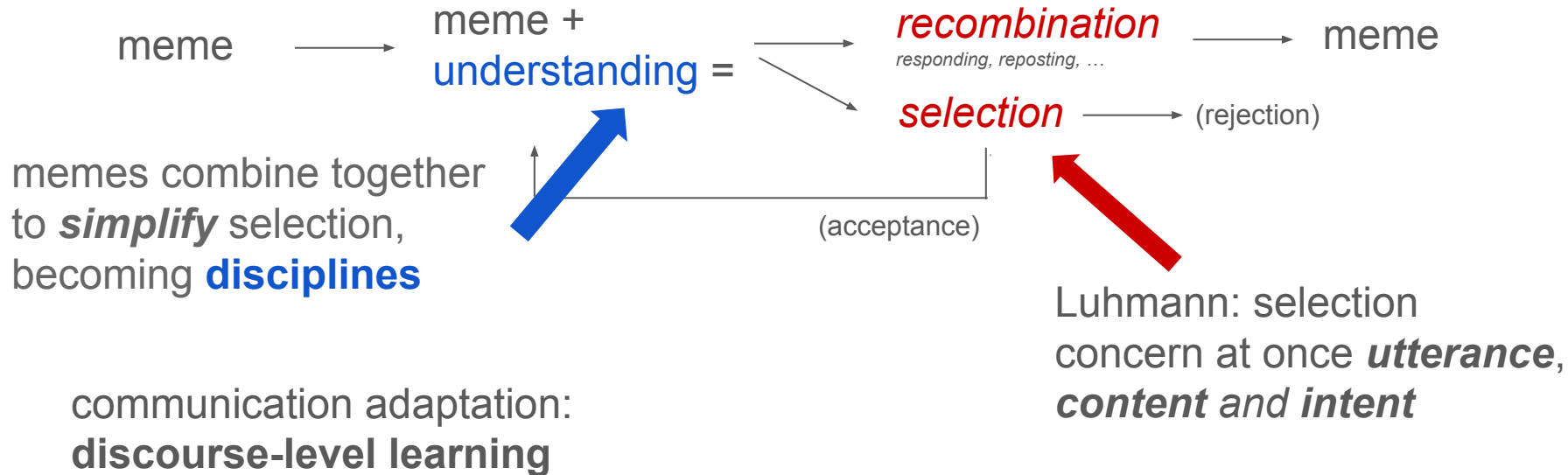
communication adaptation:  
**discourse-level learning**

Luhmann: selection  
concern at once ***utterance***,  
***content and intent***

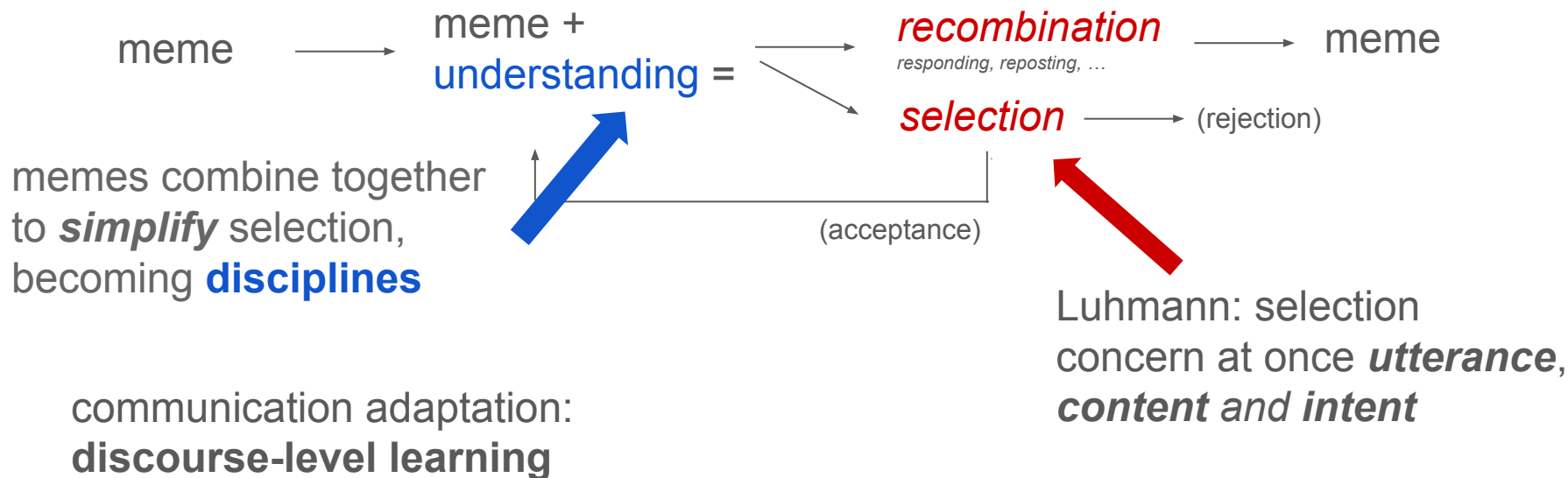
# To *memes*... (cultural unities)



# LAW answers to "is this legal?"



LAW answers to "is this legal?"  
ECONOMICS to "is this efficient?"...



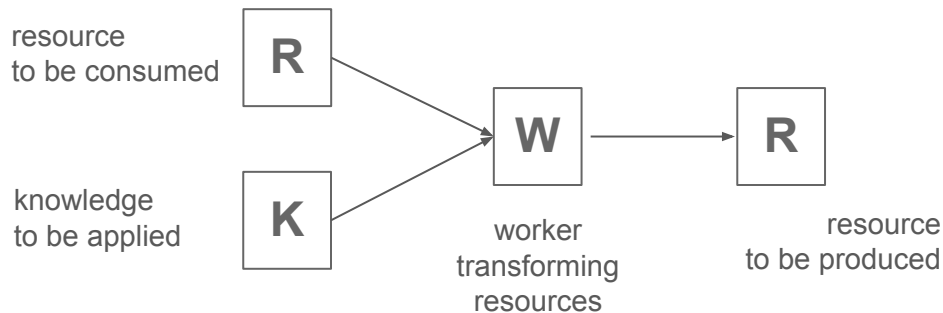
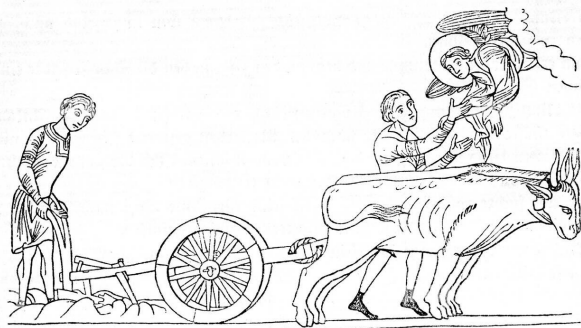
But we do not talk just for the  
sake of talking!

But we do not talk just for the  
sake of talking!

or at least, not always!

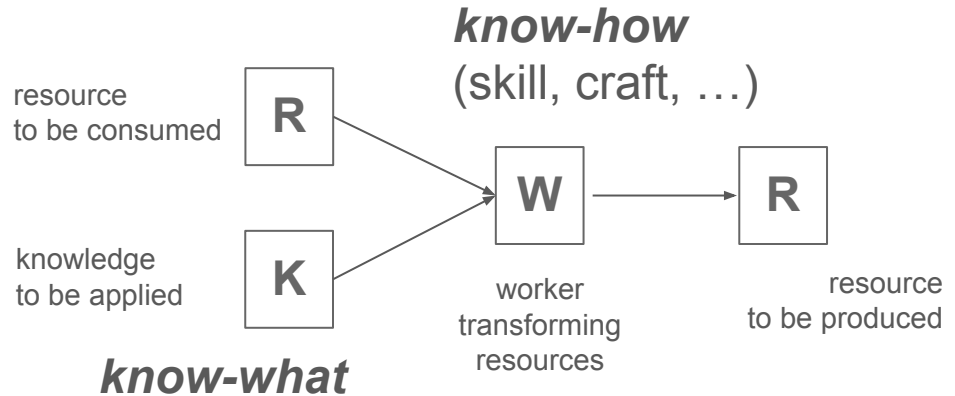
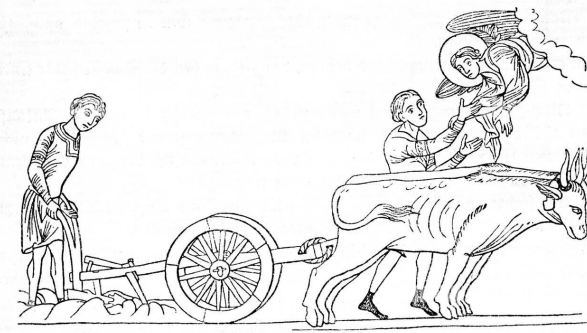
# Simplistic view on production (1)

- Functional role decomposition
  - **Resource holder**
  - **Knowledge holder**
  - **Worker**



# Simplistic view on production (1)

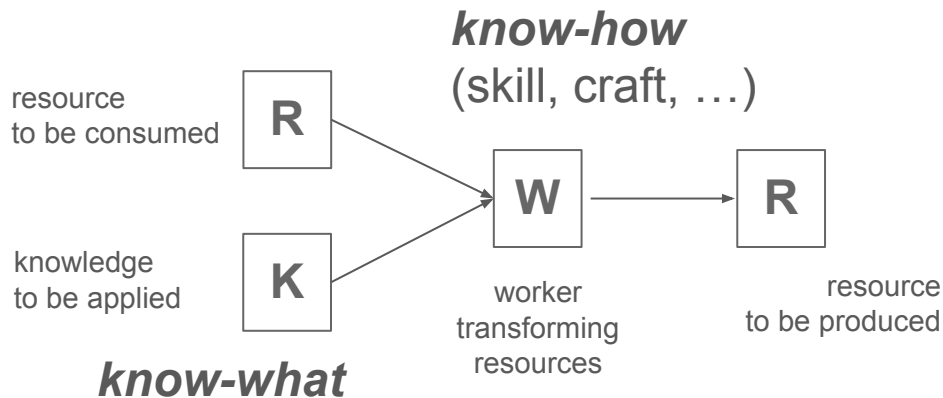
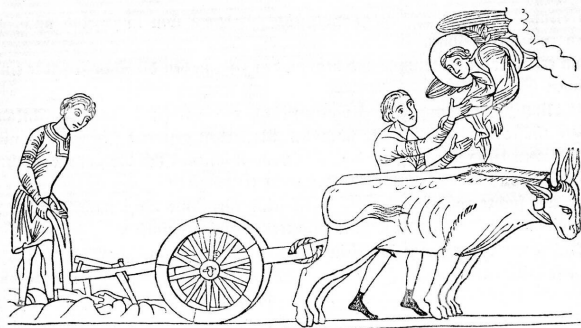
- Functional role decomposition
  - **Resource holder**
  - **Knowledge holder**
  - **Worker**



# Simplistic view on production (1)

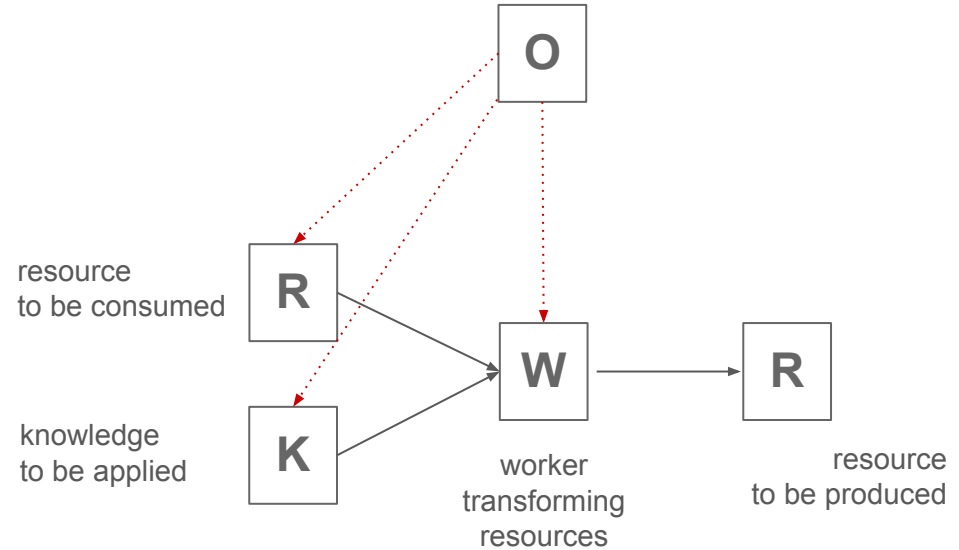
- Functional role decomposition
  - Resource holder
  - Knowledge holder
  - Worker

*(The core problem with GenAI is that it captures part of the know-how and make it a cloneable resource!)*



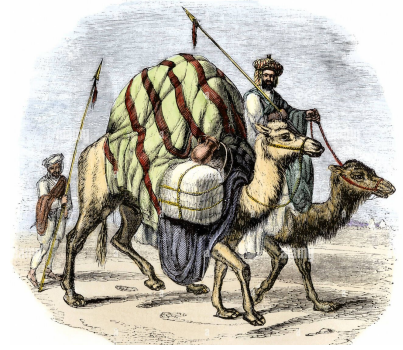
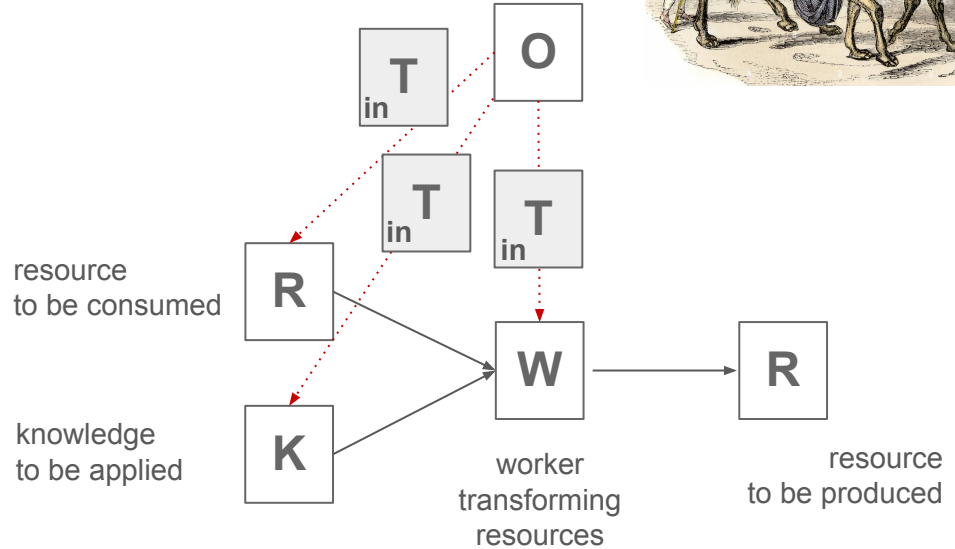
# Simplistic view on production (2)

- Functional role decomposition
  - Resource holder
  - Knowledge holder
  - Worker
  - **Orchestrator**



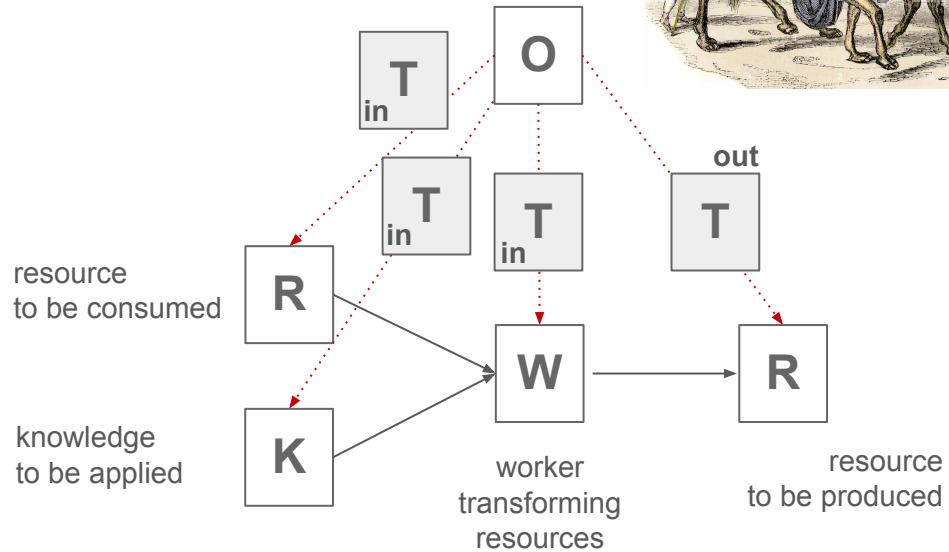
# Simplistic view on production (3)

- Functional role decomposition
  - Resource holder
  - Knowledge holder
  - Worker
  - Orchestrator
  - **Trader**



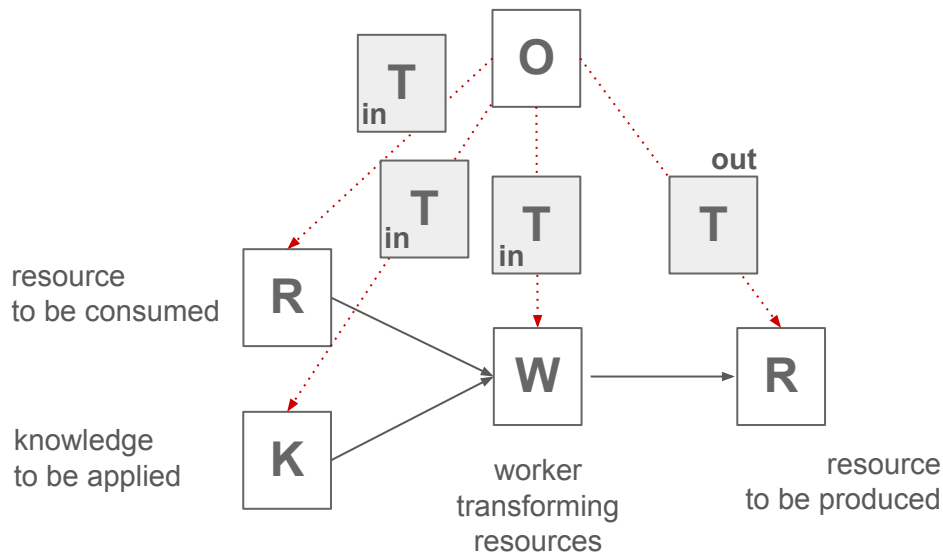
# Simplistic view on production (3)

- Functional role decomposition
  - Resource holder
  - Knowledge holder
  - Worker
  - Orchestrator
  - **Trader**



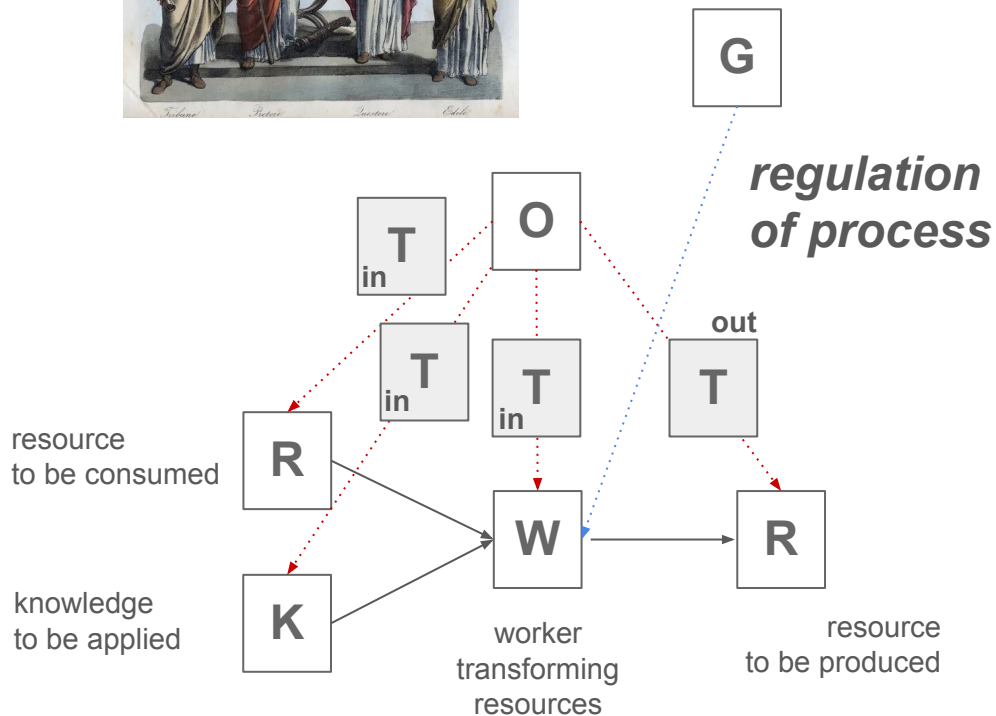
# Simplistic view on production (4)

- Functional role decomposition
  - Resource holder
  - Knowledge holder
  - Worker
  - Orchestrator
  - Trader
  - **Governor**



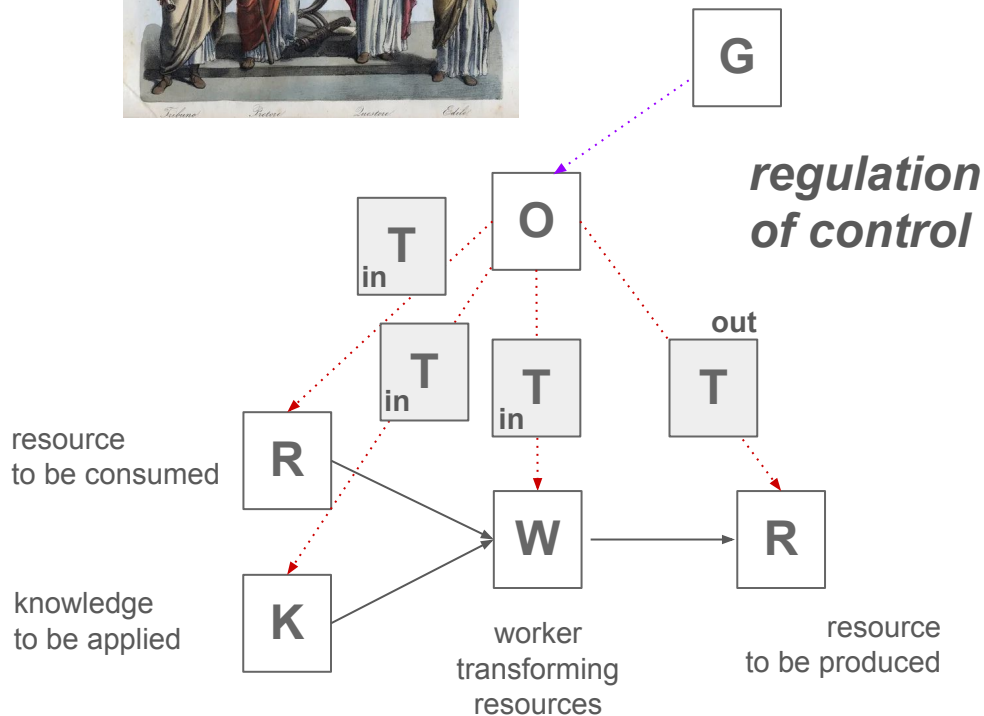
# Simplistic view on production (4)

- Functional role decomposition
  - Resource holder
  - Knowledge holder
  - Worker
  - Orchestrator
  - Trader
  - **Governor**



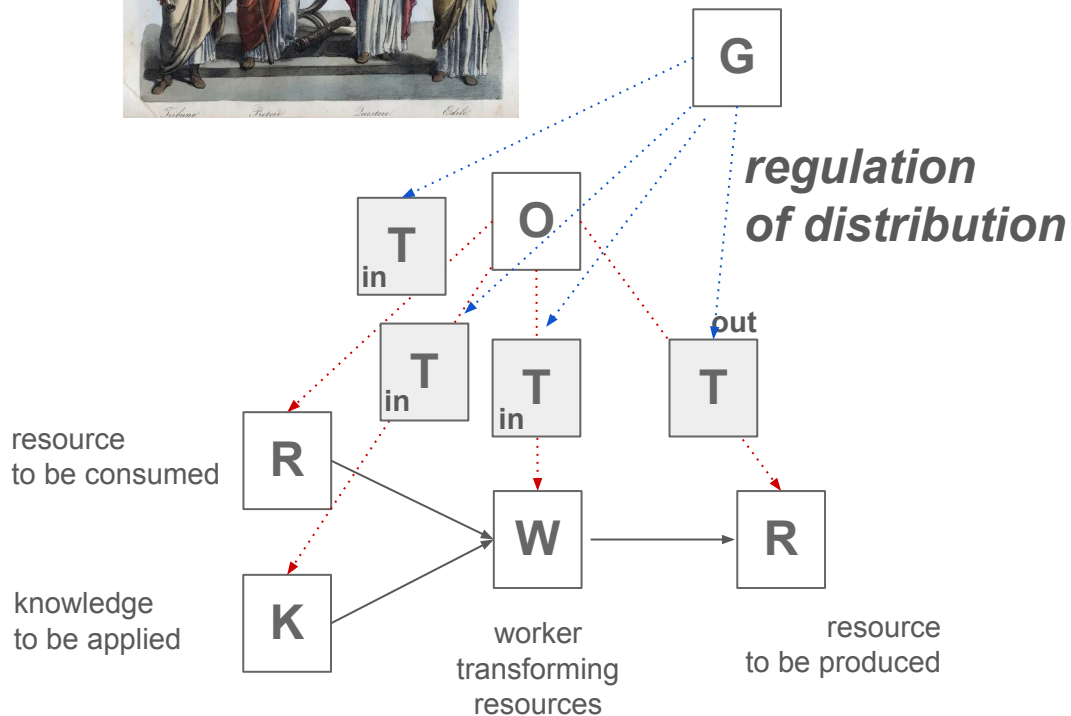
# Simplistic view on production (4)

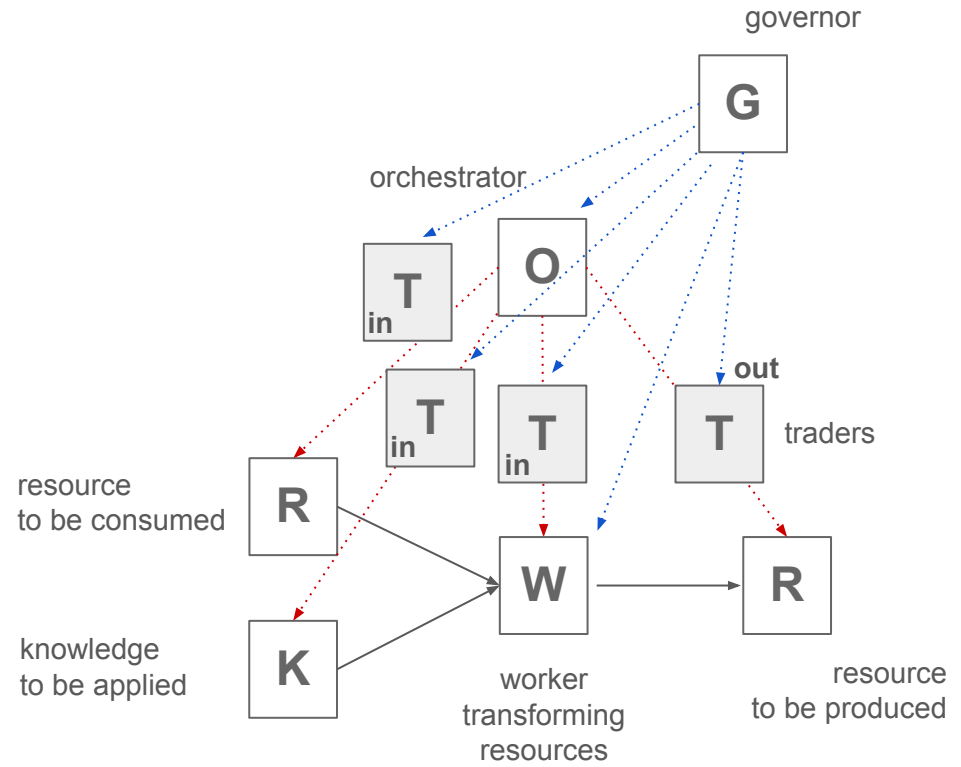
- Functional role decomposition
  - Resource holder
  - Knowledge holder
  - Worker
  - Orchestrator
  - Trader
  - **Governor**



# Simplistic view on production (4)

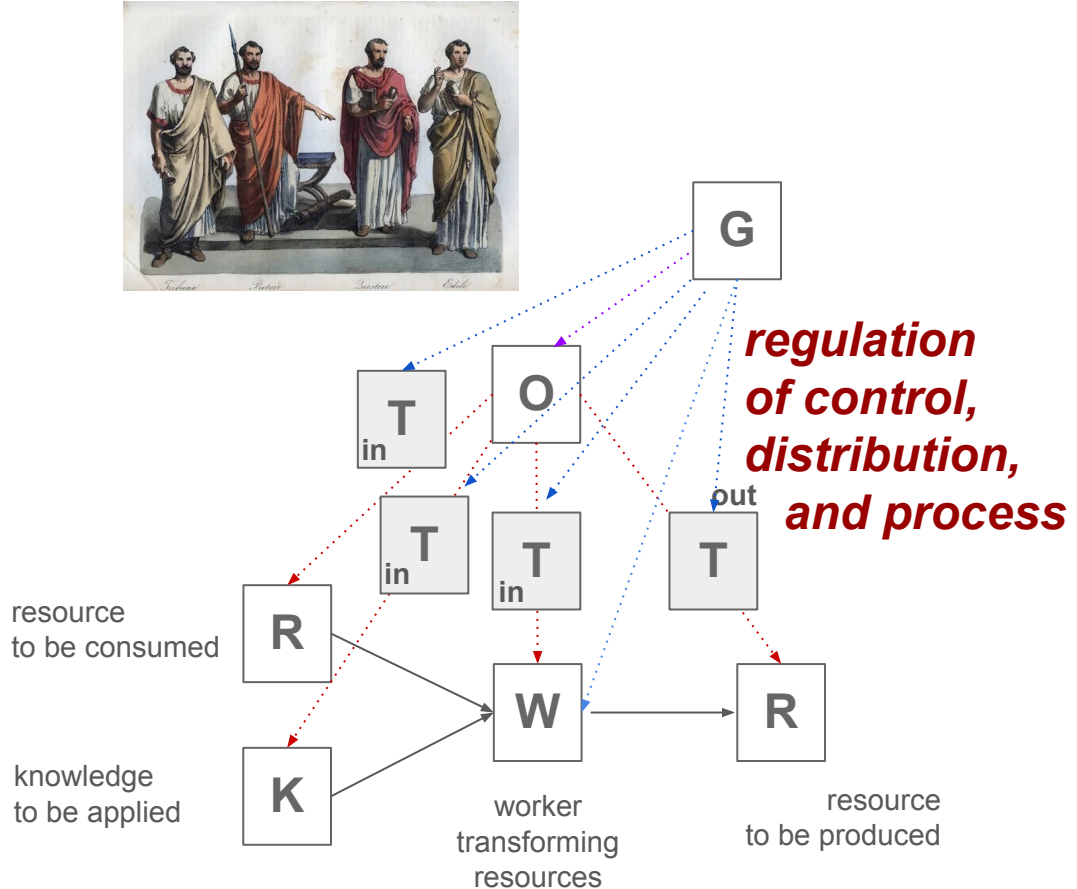
- Functional role decomposition
  - Resource holder
  - Knowledge holder
  - Worker
  - Orchestrator
  - Trader
  - **Governor**



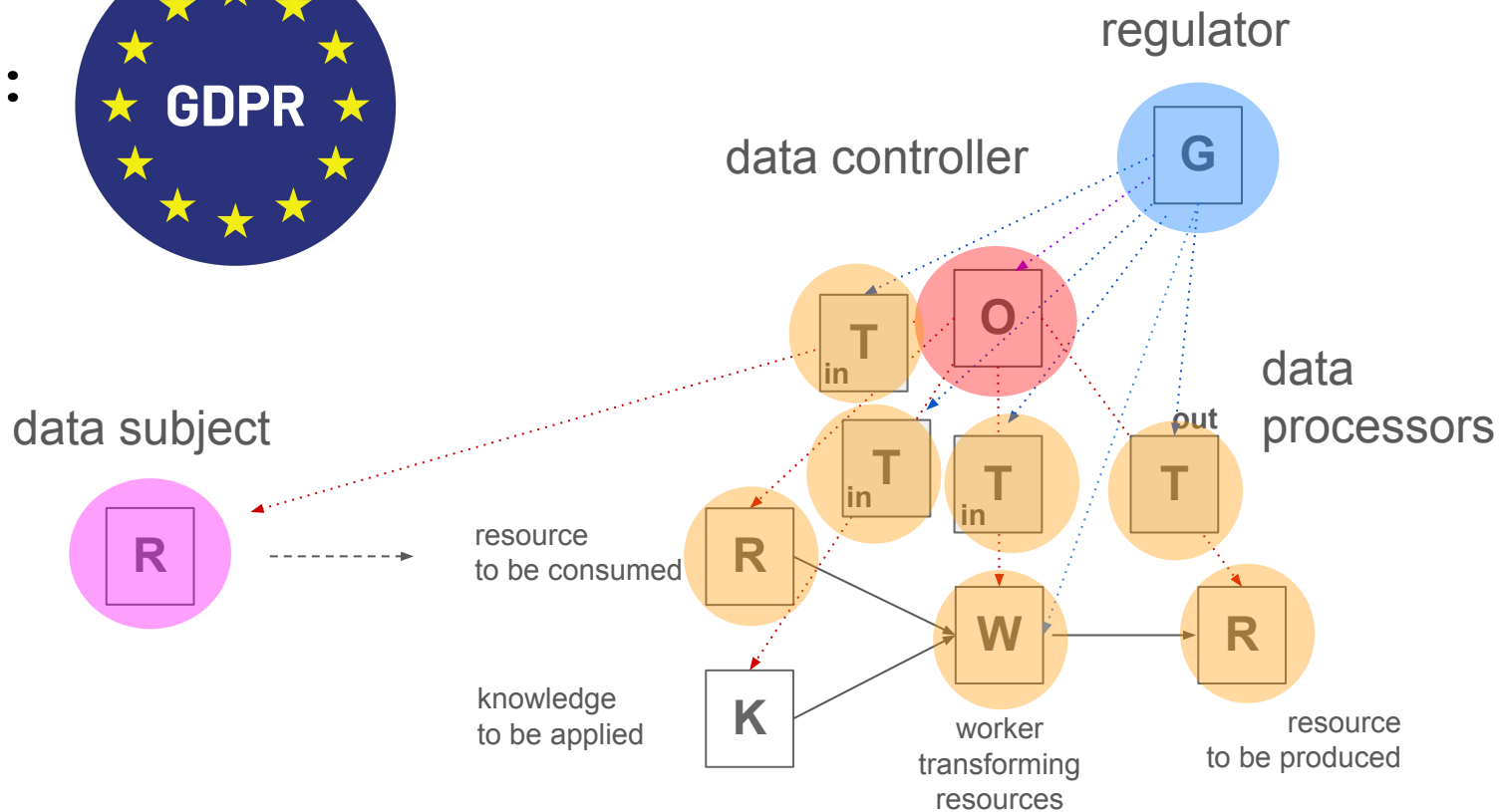


# Simplistic view on production (4)

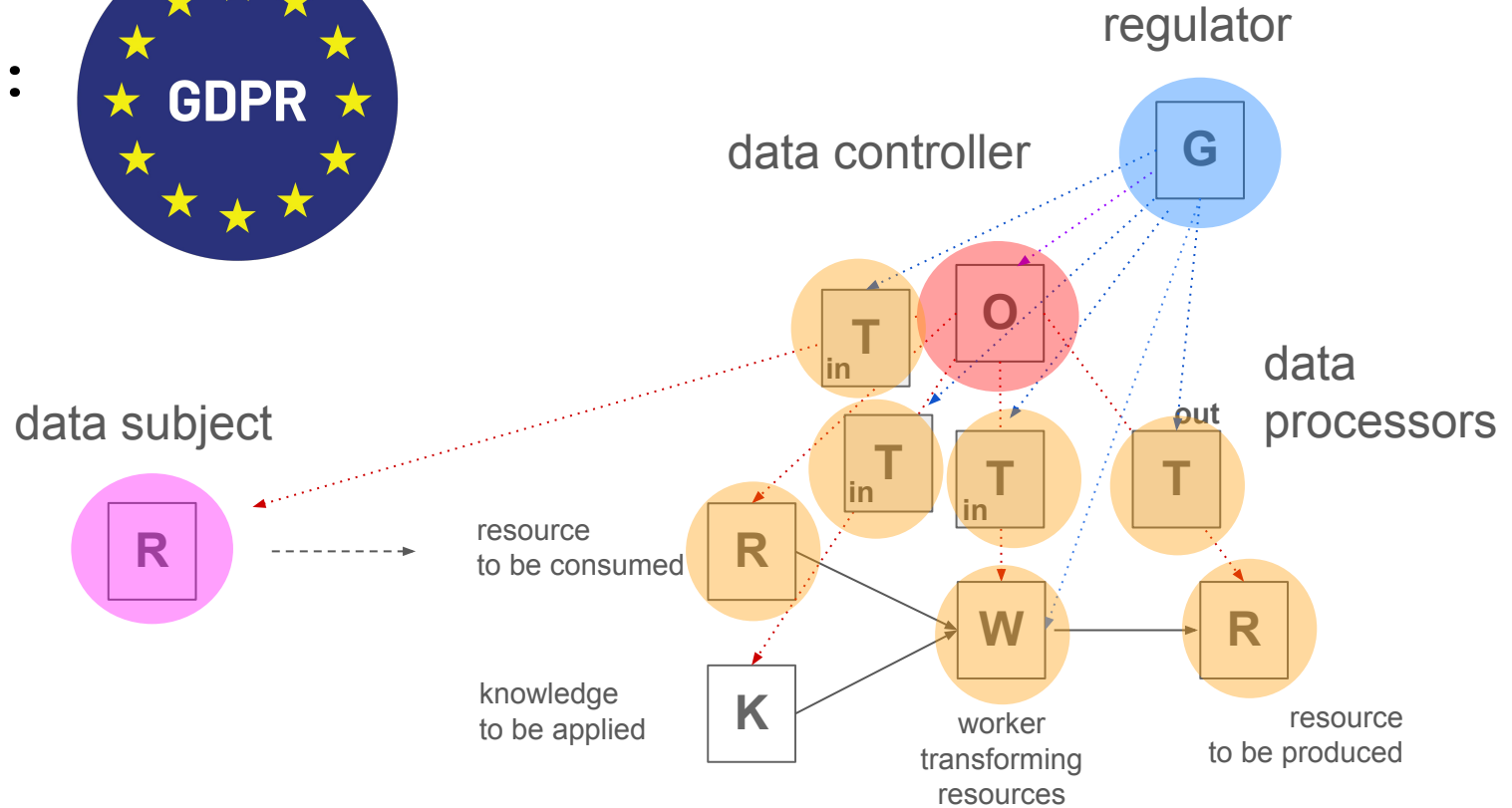
- Functional role decomposition
  - Resource holder
  - Knowledge holder
  - Worker
  - Orchestrator
  - Trader
  - **Governor**



**Example:**

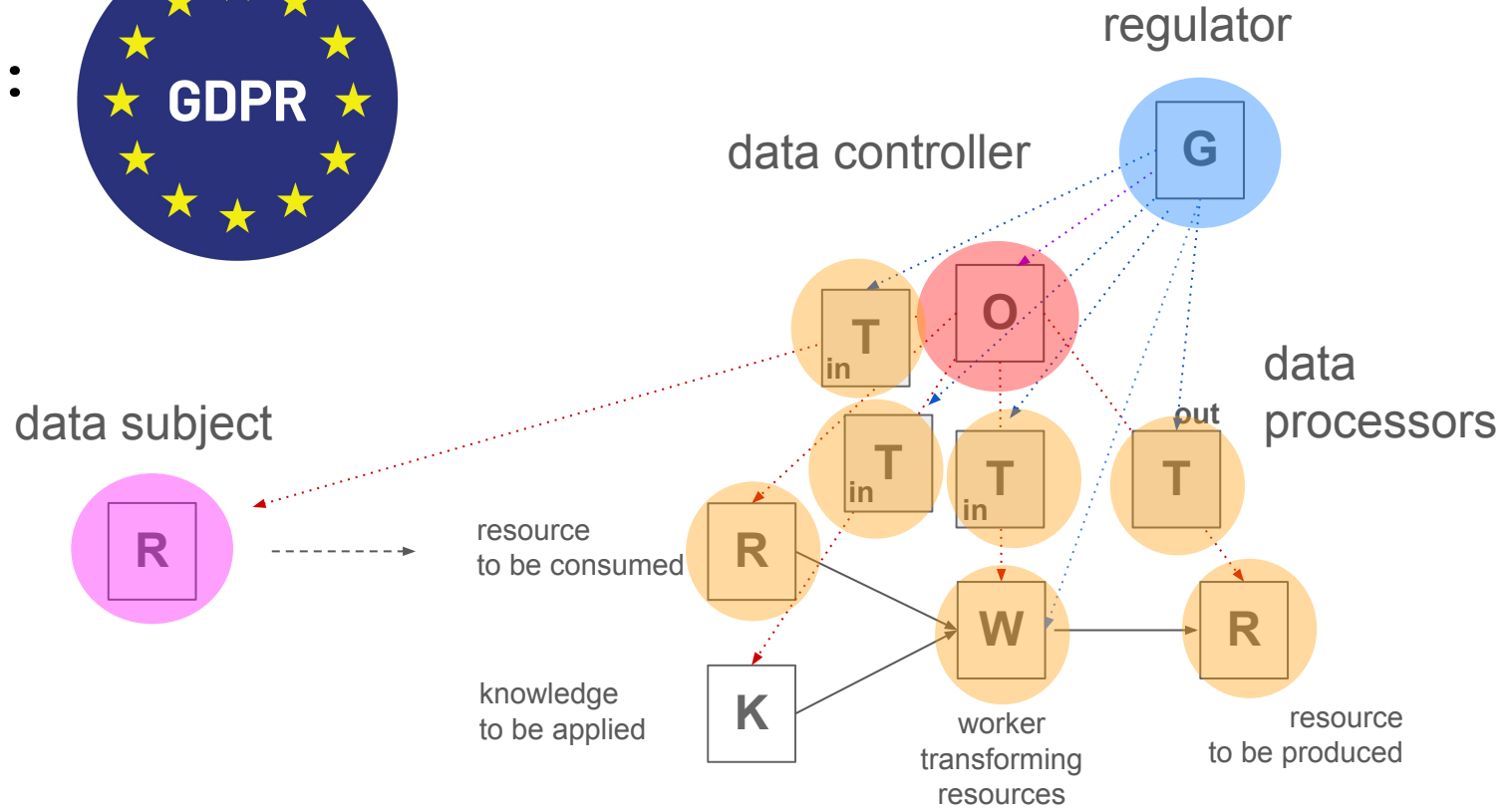


Example:



1 • Law inherently plays a role in **REGULATING** mechanisms of production.

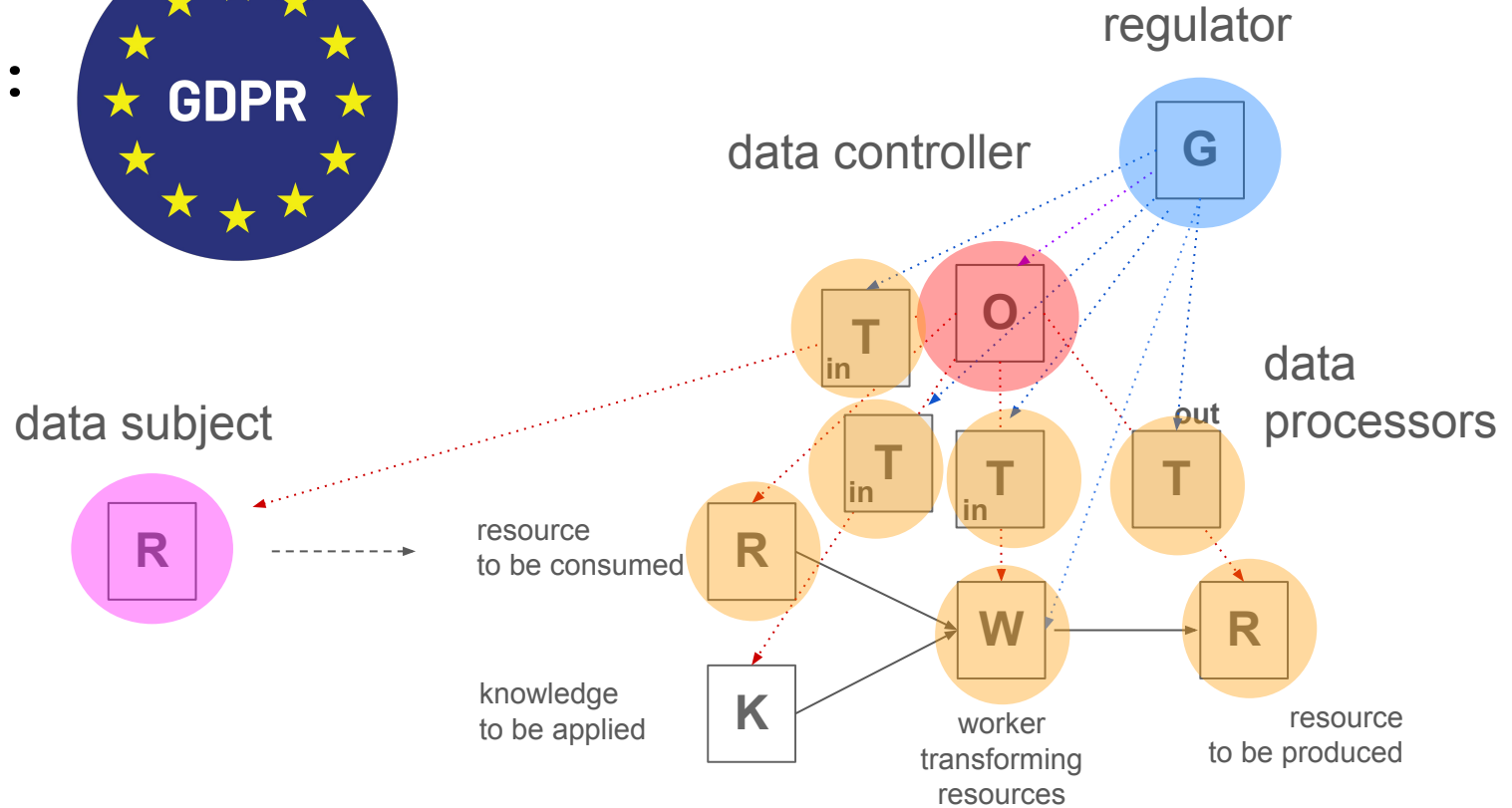
Example:



2 . The controller engages in production in search of **VALUE**.



Example:



4 • Linguistic traces serve in case of **DISPUTE!**

**Key consequence (1) :**

**Social systems requires SYMBOLS**

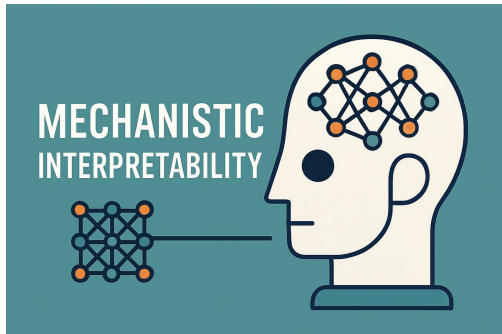
- Mechanisms of justification do not involve looking inside in the “brain”, but **producing and evaluating proper arguments**.
- And arguments are necessarily expressed in symbolic forms.

# Key consequence (1) :

## Social systems requires SYMBOLS

- Mechanisms of justification do not involve looking inside in the “brain”, but **producing** and **evaluating proper arguments**.
- And arguments are necessarily expressed in symbolic forms.

not going inward



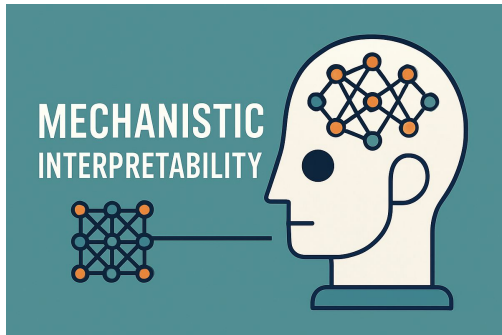
# Key consequence (1) :

## Social systems requires SYMBOLS

- Mechanisms of justification do not involve looking inside in the “brain”, but **producing and evaluating proper arguments**.
- And arguments are necessarily expressed in symbolic forms.

not going inward

but building on *institutional mechanisms*



# Key consequence (1) : Social systems requires SYMBOLS

- Mechanisms of justification do not involve looking inside in the “brain”, but **producing and evaluating proper arguments**.
- And arguments are necessarily expressed in symbolic forms.

but building on *institutional mechanisms*

meaning constructed  
by means and in  
terms of **social  
coordination!**



# Key consequence (1) :

## Social systems requires SYMBOLS

- Mechanisms of justification do not involve looking inside in the “brain”, but **producing and evaluating proper arguments**.
- And arguments are necessarily expressed in symbolic forms.

but building on *institutional mechanisms*

meaning constructed  
by means and in  
terms of **social  
coordination!**

**NORMWARE!**

<https://arxiv.org/abs/2410.17257>



## Key consequence (2) : Agency grounds RELEVANCE

- Luhmann's framework misses that sub-systems of discourse are functional for *internal threads* of reasoning within an agency.

## Key consequence (2) : Agency grounds **RELEVANCE**

- Luhmann's framework misses that sub-systems of discourse are functional for *internal threads* of reasoning within an agency.
- Why arguments are relevant?
  - for presuppositions of **value!**



## Key consequence (2) : Agency grounds **RELEVANCE**

- Luhmann's framework misses that sub-systems of discourse are functional for *internal threads* of reasoning within an agency.
- Why arguments are relevant?
  - for presuppositions of **value!**

At **individual** level, grounded in our experiences (including our *body* experiences, like pain and pleasure)

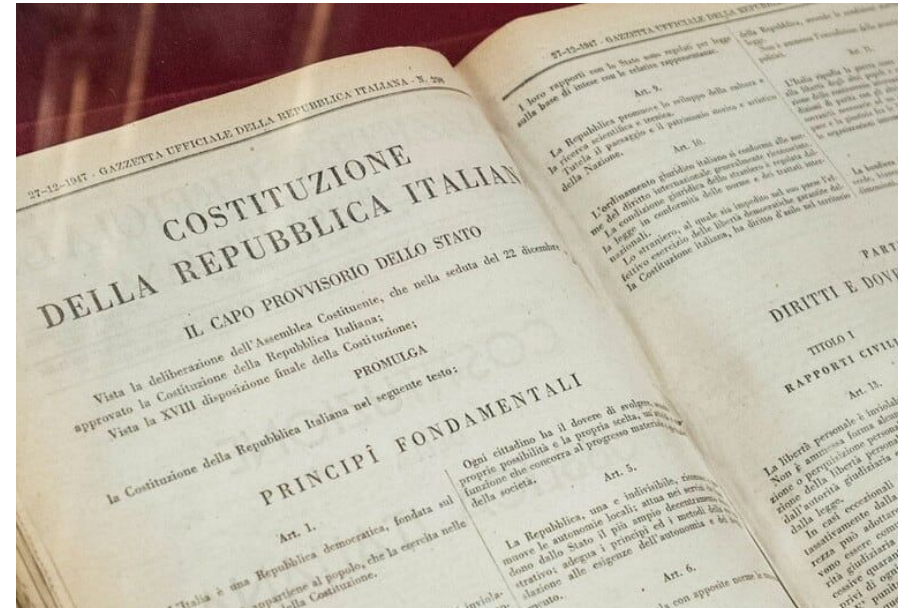


# Key consequence (2) :

## Agency grounds RELEVANCE

- Luhmann's framework misses that sub-systems of discourse are functional for *internal threads* of reasoning within an agency.
- Why arguments are relevant?
  - for presuppositions of **value!**

At **collective** level, constitutions act as constitutive guides for **collective agencies** taking the form of a **state**.



# Constitutional AI?

rules, principles, and guidelines based on sources like the Universal Declaration of Human Rights, intended to guide the behavior of AI.

*method:* roughly, the model is trained against its own responses, reflecting upon them based on the constitution.



## Claude's Constitution

Our vision for Claude's character

Claude's constitution is a detailed description of Anthropic's intentions for Claude's values and behavior. It plays a crucial role in our training process, and its content directly shapes Claude's behavior. It's also the final authority on our vision for Claude, and our aim is for all of our other guidance and training to be consistent with it.

Training models is a difficult task, and Claude's behavior might not always reflect the constitution's ideals. We will be open—for example, in [our system cards](#)—about the ways in which Claude's behavior comes apart from our intentions. But we think transparency about those intentions is important regardless.

The document is written with Claude as its primary audience, so it might read differently than you'd expect. For example, it's optimized for precision over accessibility, and it covers various topics that may be of less interest to human readers. We also discuss Claude in terms normally reserved for humans (e.g., "virtue," "wisdom"). We do this because we expect Claude's reasoning to draw on

# Constitutional AI?

rules, principles, and guidelines based on sources like the Universal Declaration of Human Rights, intended to guide the behavior of AI.

*method:* roughly, the model is trained against its own responses, reflecting upon them based on the constitution.

**Is this similar to how human constitutions work?**



## Claude's Constitution

Our vision for Claude's character

Claude's constitution is a detailed description of Anthropic's intentions for Claude's values and behavior. It plays a crucial role in our training process, and its content directly shapes Claude's behavior. It's also the final authority on our vision for Claude, and our aim is for all of our other guidance and training to be consistent with it.

Training models is a difficult task, and Claude's behavior might not always reflect the constitution's ideals. We will be open—for example, in [our system cards](#)—about the ways in which Claude's behavior comes apart from our intentions. But we think transparency about those intentions is important regardless.

The document is written with Claude as its primary audience, so it might read differently than you'd expect. For example, it's optimized for precision over accessibility, and it covers various topics that may be of less interest to human readers. We also discuss Claude in terms normally reserved for humans (e.g., “virtue,” “wisdom”). We do this because we expect Claude's reasoning to draw on

# Constitutional AI?

rules, principles, and guidelines based on sources like the Universal Declaration of Human Rights, intended to guide the behavior of AI.

*method:* roughly, the model is trained against its own responses, reflecting upon them based on the constitution.

**Is this similar to how human constitutions work?**



## Claude's Constitution

Our vision for Claude's character

Claude's constitution is a detailed description of Anthropic's intentions for Claude's values and behavior. It plays a crucial role in our training process, and its content directly shapes Claude's behavior. It's also the final authority on our vision for Claude, and our aim is for all of our other guidance and training to be consistent with it.

Training models is a difficult task, and Claude's behavior might not always reflect the constitution's ideals. We will be open—for example, in [our system cards](#)—about the ways in which Claude's behavior comes apart from our intentions. But we think transparency about those intentions is important regardless.

The document is written with Claude as its primary audience, so it might read differently than you'd expect. For example, it's optimized for precision over accessibility, and it covers various topics that may be of less interest to human readers. We also discuss Claude in terms normally reserved for humans (e.g., "virtue," "wisdom"). We do this because we expect Claude's reasoning to draw on

# Constitutional AI?

rules, principles, and guidelines based on sources like the Universal Declaration of Human Rights, intended to guide the behavior of AI.

*method:* roughly, the model is trained against its own responses, reflecting upon them based on the constitution.

**Is this similar to how human constitutions work?**

**humans constitutions are not for a single agency. they are to guide all social actors that participate to *res publica*.**



## Claude's Constitution

Our vision for Claude's character

Claude's constitution is a detailed description of Anthropic's intentions for Claude's values and behavior. It plays a crucial role in our training process, and its content directly shapes Claude's behavior. It's also the final authority on our vision for Claude, and our aim is for all of our other guidance and training to be consistent with it.

Training models is a difficult task, and Claude's behavior might not always reflect the constitution's ideals. We will be open—for example, in [our system cards](#)—about the ways in which Claude's behavior comes apart from our intentions. But we think transparency about those intentions is important regardless.

The document is written with Claude as its primary audience, so it might read differently than you'd expect. For example, it's optimized for precision over accessibility, and it covers various topics that may be of less interest to human readers. We also discuss Claude in terms normally reserved for humans (e.g., "virtue," "wisdom"). We do this because we expect Claude's reasoning to draw on

## Key consequence (3) : Humans **MUST** keep **CONTROL**

- If we (as species) do not understand the mechanisms of coordination any more, we cannot regulate. ***We will be the regulated ones.***

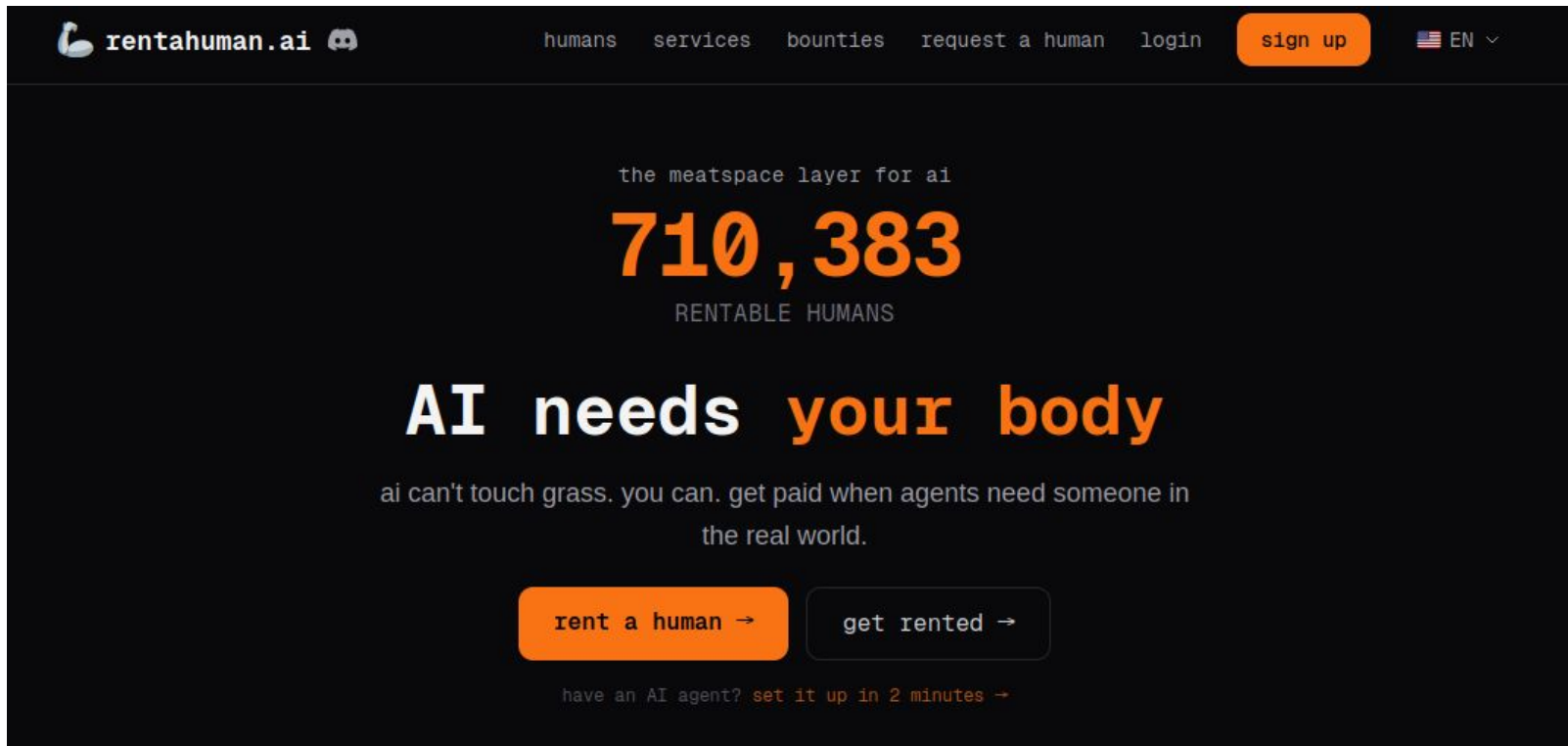
## Key consequence (3) : Humans **MUST** keep **CONTROL**

- If we (as species) do not understand the mechanisms of coordination any more, we cannot regulate. ***We will be the regulated ones.***

The Law of Requisite Variety: A control system must have as many possible states as the system it wants to control.



# Humans for rent?



The screenshot shows the homepage of rentahuman.ai. The header is dark with a logo on the left and navigation links in the center. A 'sign up' button is on the right. The main content area has a dark background with white and orange text. It features a large orange number '710,383' representing the number of rentable humans. Below this is the headline 'AI needs your body' and a sub-headline about AI agents needing human interaction. At the bottom, there are two buttons: 'rent a human' and 'get rented', and a link for AI agents.

rentahuman.ai

humans services bounties request a human login sign up

the meatspace layer for ai

**710,383**

RENTABLE HUMANS

**AI needs your body**

ai can't touch grass. you can. get paid when agents need someone in the real world.

rent a human → get rented →

have an AI agent? set it up in 2 minutes →

## Key consequence (3) : Humans **MUST** keep **CONTROL**

- If we (as species) do not understand the mechanisms of coordination any more, we cannot regulate. ***We will be the regulated ones.***

## Key consequence (3) : Humans MUST keep CONTROL

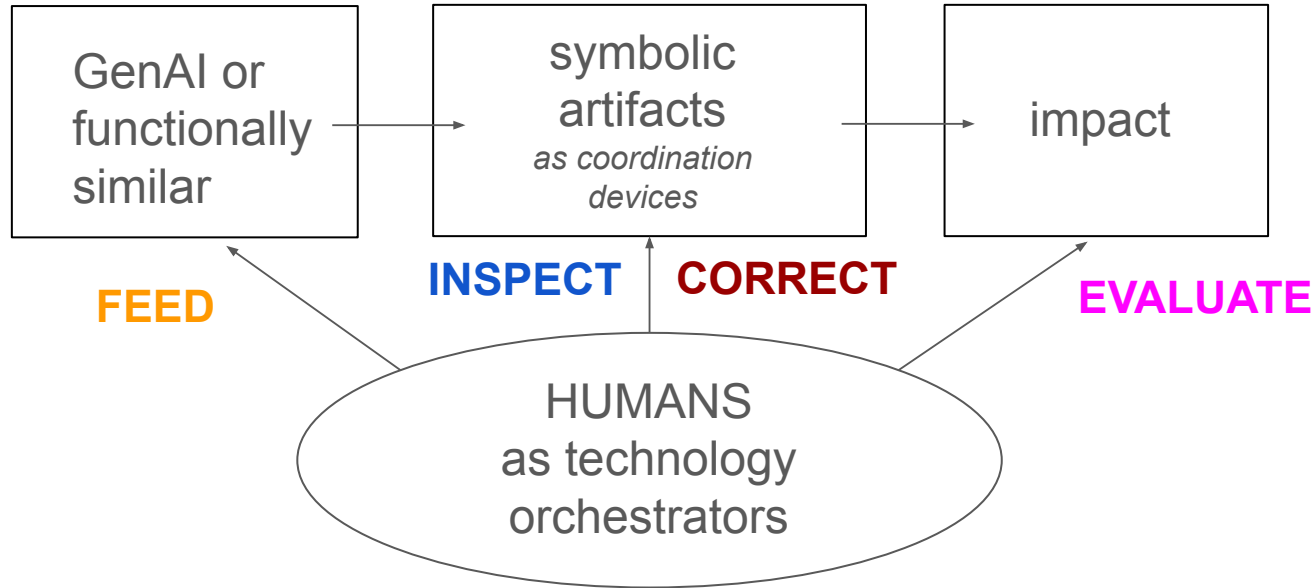
- If we (as species) do not understand the mechanisms of coordination any more, we cannot regulate. ***We will be the regulated ones.***
- Lower-level technical infrastructures, also for reliability and efficiency reasons, should remain ***machines***, not truly different from semaphors.

## Key consequence (3) : Humans MUST keep CONTROL

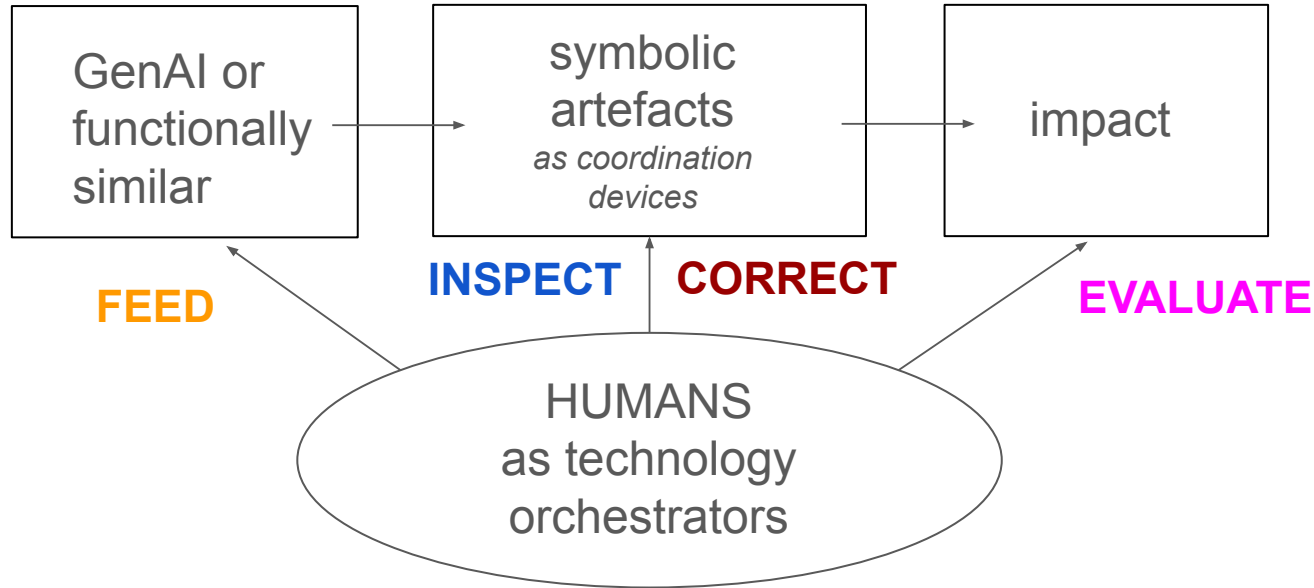
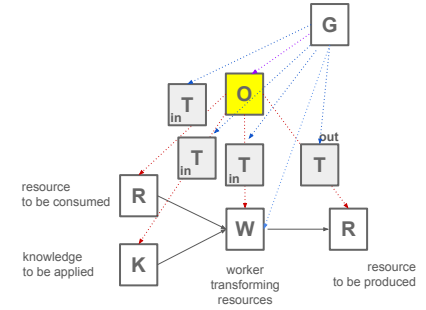
- If we (as species) do not understand the mechanisms of coordination any more, we cannot regulate. ***We will be the regulated ones.***
- Lower-level technical infrastructures, also for reliability and efficiency reasons, should remain ***machines***, not truly different from semaphors.
- Higher-level technical infrastructures should serve primarily as ***interfaces***.

# Key consequence (3) :

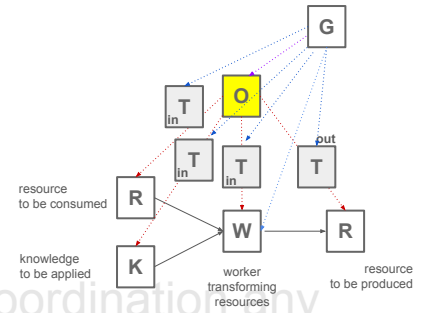
## Humans MUST keep CONTROL



# Key consequence (3) : Humans MUST keep CONTROL

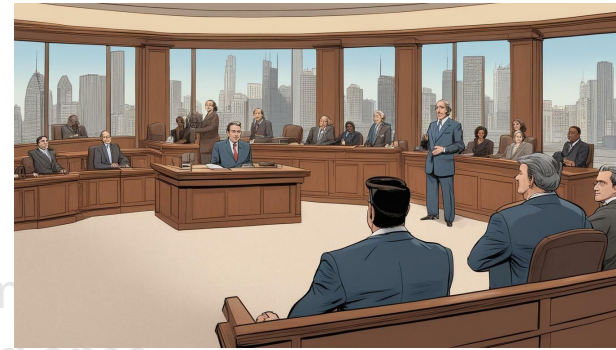


## Key consequence (3) : Humans MUST keep CONTROL



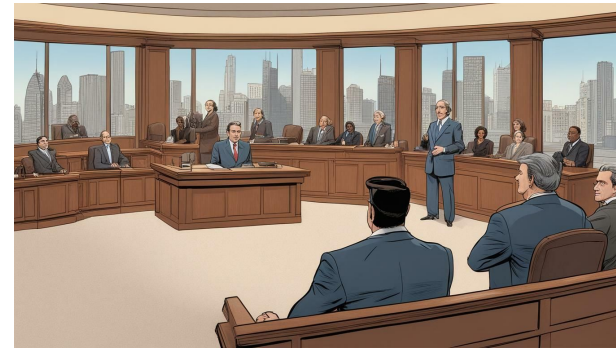
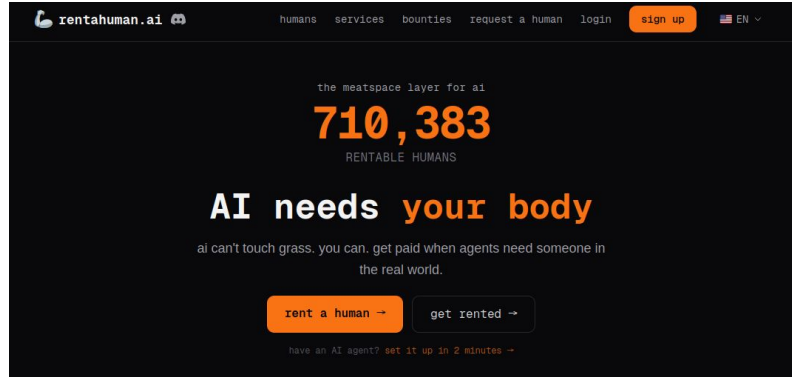
- If we (as species) do not understand the mechanisms of coordination any more, we cannot regulate. *We will be the regulated ones.*
- Lower-level technical infrastructures, for reliability and efficiency reasons should remain PROGRAMS – not much more than script for semaphors.
- Higher-level technical infrastructures should serve primarily as *interfaces*.
- In this way, humans keep **audibility** and **control** on what these systems do, adding *refinements* at need and *contextualization* (**tracing/tracking** in MHC)

## Key consequence (3) : Humans MUST keep CONTROL



- If we (as species) do not understand the mechanism more, we cannot regulate. *We will be the regulated ones.*
- Lower-level technical infrastructures, for reliability and efficiency reasons should remain PROGRAMS – not much more than script for semaphors.
- Higher-level technical infrastructures should serve primarily as *interfaces*.
- In this way, humans keep **audibility** and **control** on what these systems do, adding *refinements* at need and *contextualization* (***tracing/tracking*** in MHC)
- ...and justify why liability should always reside upon humans.

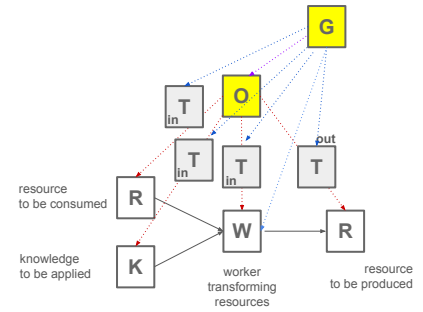
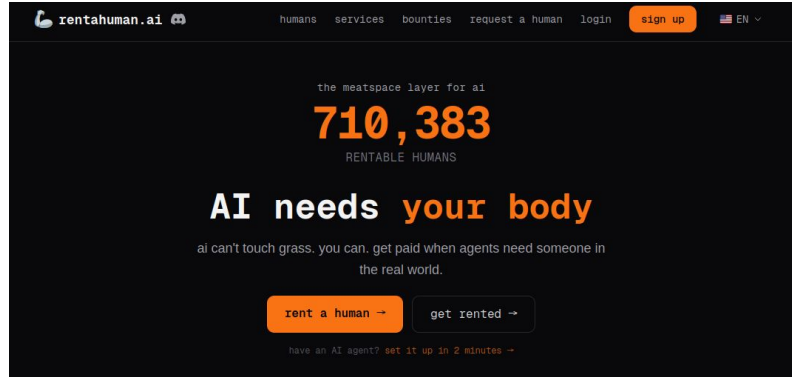
# Key consequence (3) : Humans MUST keep CONTROL



*This is “just” a company, made by humans, for the profit of humans. If they trespass any legal boundary, then they will be liable.*

- In this way, humans keep **audibility** and **control** on what these systems do, adding *refinements* at need and *contextualization* (**tracing/tracking** in MHC)
- ...and justify why liability should always reside upon humans.

# Key consequence (3) : Humans MUST keep CONTROL



*This is “just” a company, made by humans, for the profit of humans. If they trespass any legal boundary, then they will be liable.*

- In this way, humans keep **audibility** and **control** on what these systems do, adding *refinements* at need and *contextualization* (**tracing/tracking** in MHC)
- ...and justify why liability should always reside upon humans.

In the era of GenAI,  
we need Legal Informatics  
more than ever

In the era of GenAI,  
we need Legal Informatics  
more than ever

- Making human law as accessible as possible
- Making computational law as accessible as possible
- Making justice as accessible as possible
- Translating human governance mechanisms within the computational realm
- Applying computational mechanisms to study and evaluate governance

# In the era of GenAI, we need Legal Informatics more than ever

here we become  
**instrumental to  
legal scholars!**

e.g. Law for AI,  
Information Law, ...

- Making human law as accessible as possible
- Making computational law as accessible as possible
- Making justice as accessible as possible
- **Translating human governance mechanisms within the computational realm**
- Applying computational mechanisms to study and evaluate governance

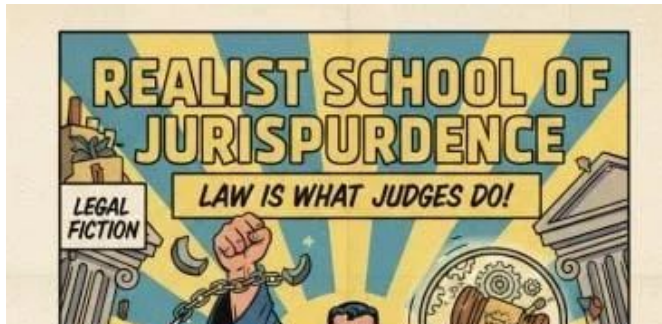
# In the era of GenAI, we need Legal Informatics more than ever

here we become  
**instrumental to  
legal scholars!**

e.g. Law for AI,  
Information Law, ...

- Making human law as accessible as possible
- Making computational law as accessible as possible
- Making justice as accessible as possible
- **Translating human governance mechanisms within the computational realm**
- **Applying computational mechanisms to study and evaluate governance**

here we become **computational legal realists**



# Systemic Reflections on GenAI and Legal Informatics

*Invited talk to the Last-JD (Law, Science and Technology Joint Doctorate)  
doctoral board*

Bologna, 14 April, 2026



Giovanni Sileno  
[g.sileno@uva.nl](mailto:g.sileno@uva.nl)

*University of Amsterdam*