How'd We Arrive Here?

(Selmer's Leibnizian Whirlwind History of Logic, with Discussion of "The Singularity")

Selmer Bringsjord

Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Lally School of Management & Technology
Rensselaer Polytechnic Institute (RPI)
Troy, New York 12180 USA

Intro to Logic
1/23/2023

R A I R
Rensselaer AI and Reasoning Lab

Logic-and-AI in the news

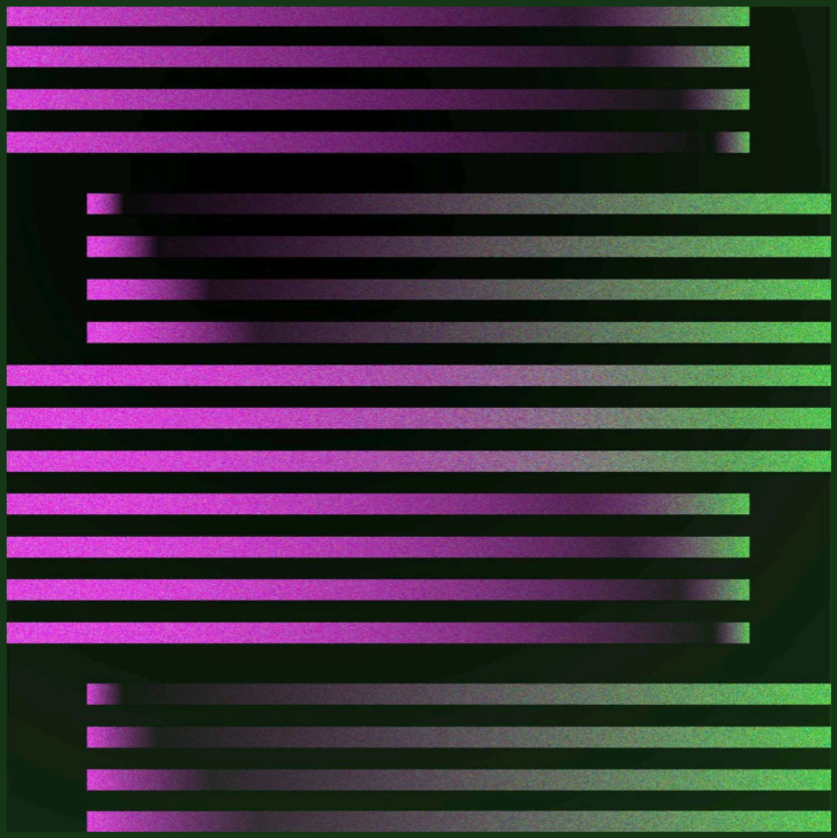
…

OpenAI API RESEARCH BLOG ABOUT

ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

TRY CHATGPT ↗



November 30, 2022
13 minute read

“These principles are often derived from a combination of different ethical theories and perspectives, such as consequentialism, deontology, virtue ethics, and care ethics.”

“These principles are often derived from a combination of different ethical theories and perspectives, such as consequentialism, deontology, virtue ethics, and care ethics.”

“he ethical principles and values that guide the development and use of AI and language models, such as transparency, fairness, non-discrimination, and privacy, are …”

“These principles are often derived from a combination of different ethical theories and perspectives, such as consequentialism, deontology, virtue ethics, and care ethics.”

“he ethical principles and values that guide the development and use of AI and language models, such as transparency, fairness, non-discrimination, and privacy, are …”

arXiv:2203.02155v1 [cs.CL] 4 Mar 2022

Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*
Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray
John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens
Amanda Askell† Peter Welinder Paul Christiano*†
Jan Leike* Ryan Lowe*

OpenAI

Abstract

Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not *aligned* with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through the OpenAI API, we collect a dataset of labeler demonstrations of the desired model behavior, which we use to fine-tune GPT-3 using supervised learning. We then collect a dataset of rankings of model outputs, which we use to further fine-tune this supervised model using reinforcement learning from human feedback. We call the resulting models *InstructGPT*. In human evaluations on our prompt distribution, outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3, despite having 100x fewer parameters. Moreover, InstructGPT models show improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets. Even though InstructGPT still makes simple mistakes, our results show that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

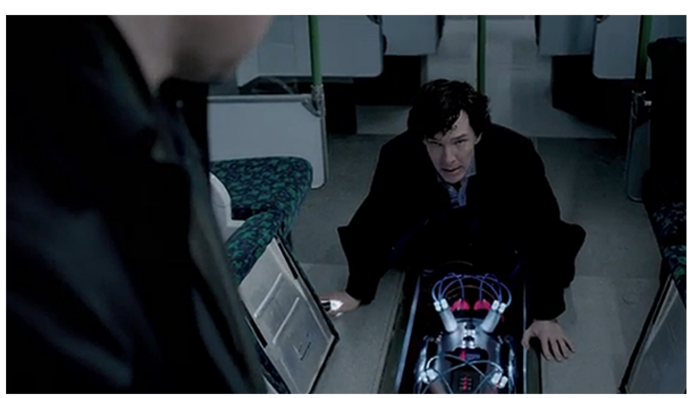
1 Introduction

Large language models (LMs) can be “prompted” to perform a range of natural language processing (NLP) tasks, given some examples of the task as input. However, these models often express unintended behaviors such as making up facts, generating biased or toxic text, or simply not following user instructions (Bender et al., 2021; Bommasani et al., 2021; Kenton et al., 2021; Weidinger et al., 2021; Tamkin et al., 2021; Gehman et al., 2020). This is because the language modeling objective

*Primary authors. This was a joint project of the OpenAI Alignment team. RL and JL are the team leads. Corresponding author: lowe@openai.com.
†Work done while at OpenAI. Current affiliations: AA: Anthropic; PC: Alignment Research Center.

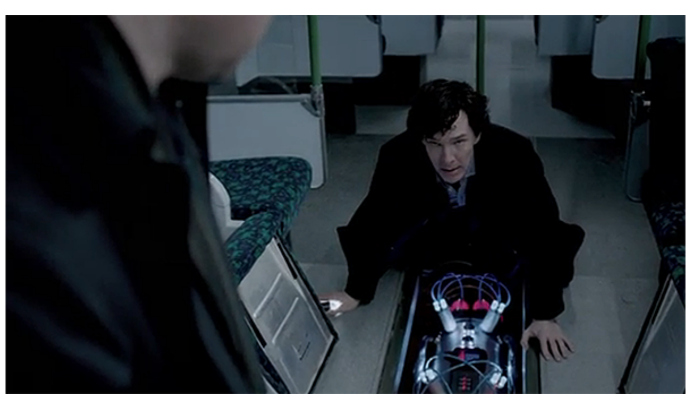
Last time …

A criminal genius nearly a match for Sherlock Holmes (Do you recognize the Dr?) has built a massive hydrogen bomb, and life on Earth is hanging in the balance, hinging on whether you make the logical prediction. Dr M gives you a sporting chance to: make the right prediction, snip or not snip accordingly, and prove that you're right …



A criminal genius nearly a match for Sherlock Holmes
(Do you recognize the Dr?)

A criminal genius nearly a match for Sherlock Holmes (Do you recognize the Dr?) has built a massive hydrogen bomb, and life on Earth is hanging in the balance, hinging on whether you make the logical prediction. Dr M gives you a sporting chance to: make the right prediction, snip or not snip accordingly, and prove that you're right …



If one of the following assertions is true then so is the other:

(1) If the red wire runs to the bomb, then the blue wire runs to the bomb; and, if the blue wire runs to the bomb, then the red wire runs to the bomb.

(2) The red wire runs to the bomb.

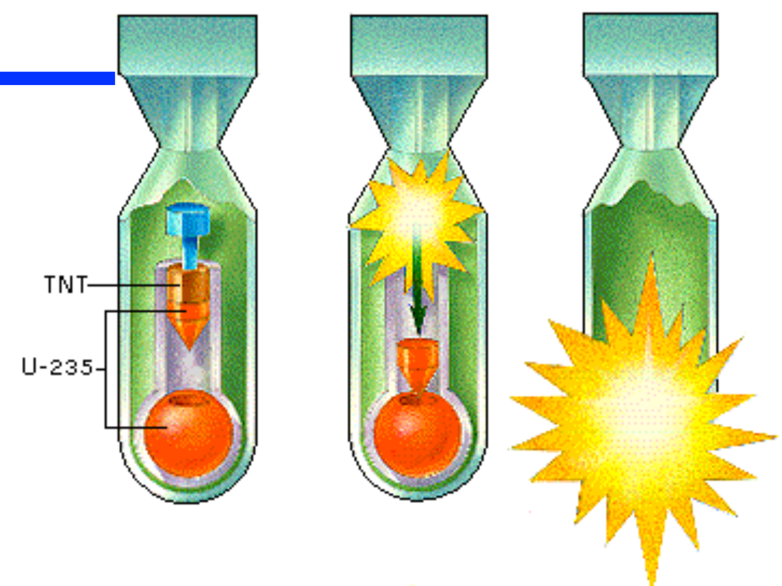
Given this perfectly reliable clue from Dr Moriarty, if either wire is more likely to run to the bomb, that wire *does* run to the bomb, and the bomb is ticking, with only a minute left! If both are equiprobable, neither runs to the bomb, and you are powerless. Make your prediction as to what will happen when a wire is snipped, and then make your selected snip by clicking on the wire you want to snip! Or leave well enough alone!

<u>Red more likely.</u>

Blue more likely.

Equiprobable.

Snip



Life on Earth has ended .

advance one more slide to see a proof that you indeed made an irrational decision…

Proposition: The blue wire is more likely!

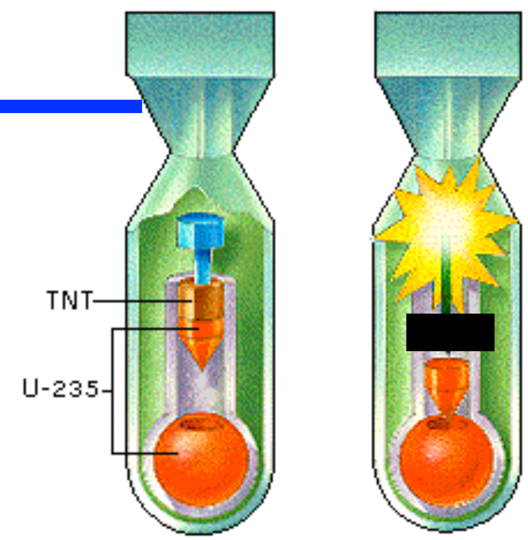
Proof: (1) can be treated as a biconditional, obviously (R <=> B).

There are two top-level cases to consider: (1) and (2) are both true; or both are false. In the case where they are both true, it's trivial to deduce both R and B. So far, then, R and B are equiprobable. What happens in the case where (1) and (2) are both false? We immediately have ~R from the denial of (2). But a biconditional is true just in case both sides are true, or both sides are false; so we have two sub-cases to consider.

Consider first the case where R is true and B is false. We have an immediate contradiction in this sub-case, so both R and B can both be deduced here, and we have not yet departed from equiprobable. So what about the case where R is false and B is true? The falsity of R is not new information (we already have that from the denial of (2)), but we can still derive B. Hence the blue wire is more likely. **QED**

STOP

Snip



Life on Earth is saved!

if you can now hand Dr M a proof that your decision was the rational one!

Advance one more slide to see a proof from Bringsjord that yours had better match up to …

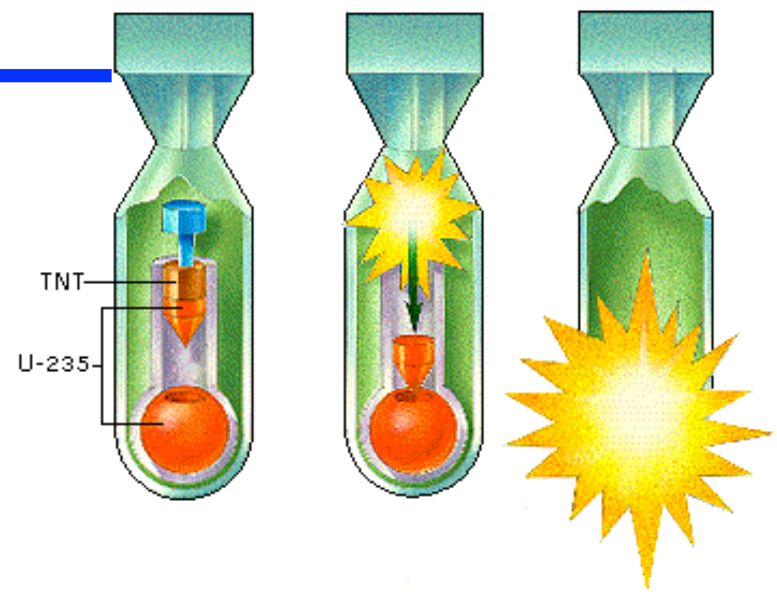
Proposition: The blue wire is more likely!

Proof: (1) can be treated as a biconditional, obviously (R <=> B).

There are two top-level cases to consider: (1) and (2) are both true; or both are false. In the case where they are both true, it's trivial to deduce both R and B. So far, then, R and B are equiprobable. What happens in the case where (1) and (2) are both false? We immediately have ~R from the denial of (2). But a biconditional is true just in case both sides are true, or both sides are false; so we have two sub-cases to consider.

Consider first the case where R is true and B is false. We have an immediate contradiction in this sub-case, so both R and B can both be deduced here, and we have not yet departed from equiprobable. So what about the case where R is false and B is true? The falsity of R is not new information (we already have that from the denial of (2)), but we can still derive B. Hence the blue wire is more likely. **QED**

STOP



Life on Earth has ended.

advance one more slide to see a proof that you indeed made an irrational decision…

Proposition: The blue wire is more likely!

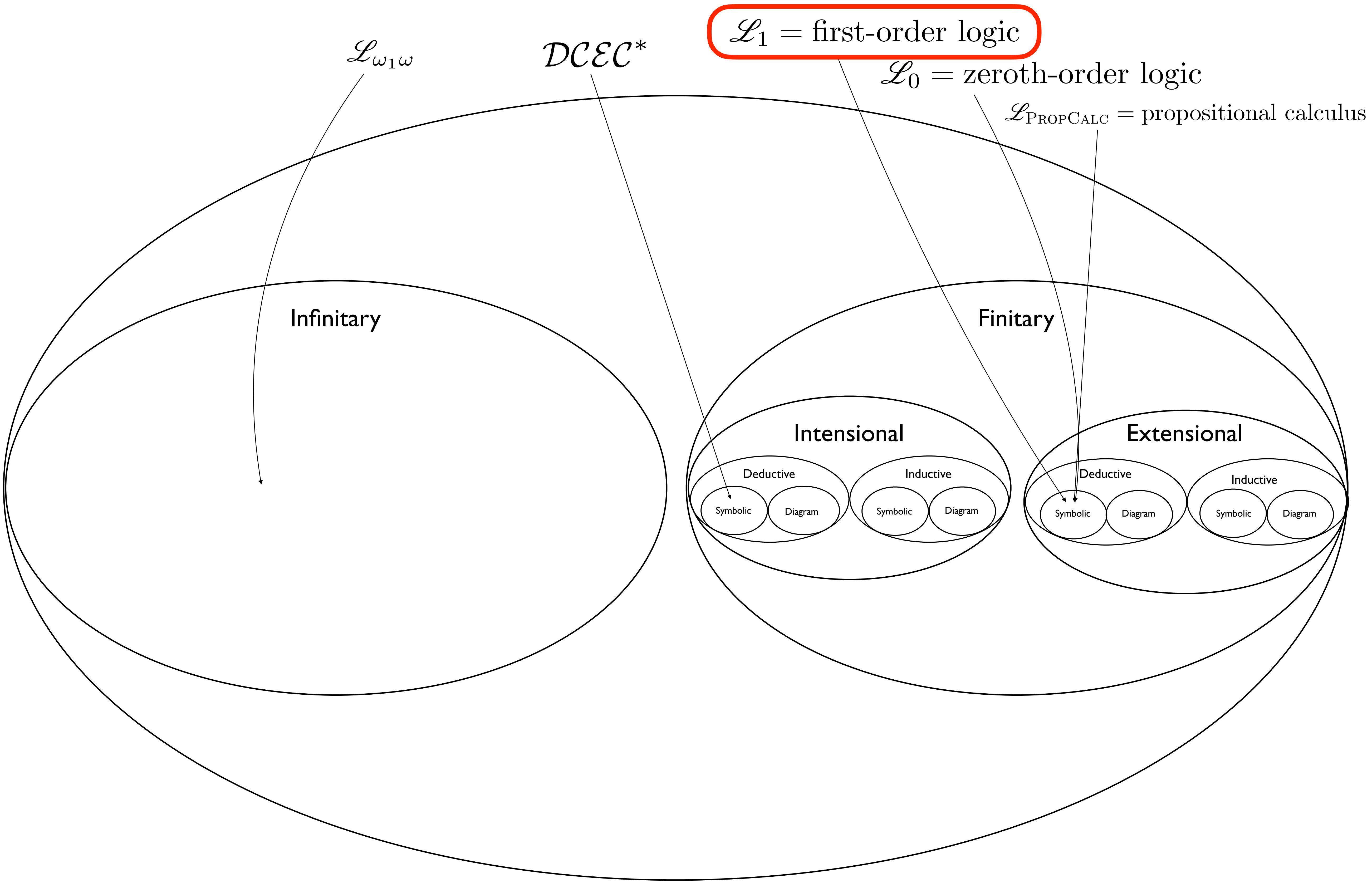
Proof: (1) can be treated as a biconditional, obviously (R <=> B).

There are two top-level cases to consider: (1) and (2) are both true; or both are false. In the case where they are both true, it's trivial to deduce both R and B. So far, then, R and B are equiprobable. What happens in the case where (1) and (2) are both false? We immediately have ~R from the denial of (2). But a biconditional is true just in case both sides are true, or both sides are false; so we have two sub-cases to consider.

Consider first the case where R is true and B is false. We have an immediate contradiction in this sub-case, so both R and B can both be deduced here, and we have not yet departed from equiprobable. So what about the case where R is false and B is true? The falsity of R is not new information (we already have that from the denial of (2)), but we can still derive B. Hence the blue wire is more likely. **QED**

STOP

The Universe of Logics



Special Llamas Disjunction

There's a thing such that it's both a llama and a non-llama;
or
there's a thing such that if it's a llama, everything is a llama;
or
there's a thing such that every llama is a non-llama.

Special Llamas Disjunction

There's a thing such that it's both a llama and a non-llama;
or
there's a thing such that if it's a llama, everything is a llama;
or
there's a thing such that every llama is a non-llama.

Is this disjunction TRUE, FALSE, or UNKNOWN?

Special Llamas Disjunction

There's a thing such that it's both a llama and a non-llama;
or
there's a thing such that if it's a llama, everything is a llama;
or
there's a thing such that every llama is a non-llama.

Is this disjunction TRUE, FALSE, or UNKNOWN?

Special Llamas Disjunction

There's a thing such that it's both a llama and a non-llama;
or
there's a thing such that if it's a llama, everything is a llama;
or
there's a thing such that every llama is a non-llama.

Is this disjunction TRUE, FALSE, or UNKNOWN?

Special Llamas Disjunction

There's a thing such that it's both a llama and a non-llama;
or
there's a thing such that if it's a llama, everything is a llama;
or
there's a thing such that every llama is a non-llama.

Is this disjunction TRUE, FALSE, or UNKNOWN?

Supply a formal proof!

Background Claim

$\mathcal{R}$ Humans, at least neurobiologically normal ones, are fundamentally rational, where rationality is constituted by certain logico-mathematically based reasoning and decision-making in response to real-world stimuli, including stimuli given in the form of focused tests; but mere animals are not fundamentally rational, since, *contra* Darwin, their minds are fundamentally qualitatively inferior to the human mind. As to whether computing machines/robots are fundamentally rational, the answer is "No." For starters, if x can't read, write, and create, x can't be rational; computing machines/robots can neither read nor write nor create; ergo, they aren't fundamentally rational.

abstract-and-valid inference schemata

Background Claim

$\mathcal{R}$ Humans, at least neurobiologically normal ones, are fundamentally rational, where rationality is constituted by certain logico-mathematically based reasoning and decision-making in response to real-world stimuli, including stimuli given in the form of focused tests; but mere animals are not fundamentally rational, since, *contra* Darwin, their minds are fundamentally qualitatively inferior to the human mind. As to whether computing machines/robots are fundamentally rational, the answer is "No." For starters, if x can't read, write, and create, x can't be rational; computing machines/robots can neither read nor write nor create; ergo, they aren't fundamentally rational.

abstract-and-valid inference schemata

quantification

Background Claim

$\mathcal{R}$ Humans, at least neurobiologically normal ones, are fundamentally rational, where rationality is constituted by certain logico-mathematically based reasoning and decision-making in response to real-world stimuli, including stimuli given in the form of focused tests; but mere animals are not fundamentally rational, since, *contra* Darwin, their minds are fundamentally qualitatively inferior to the human mind. As to whether computing machines/robots are fundamentally rational, the answer is "No." For starters, if x can't read, write, and create, x can't be rational; computing machines/robots can neither read nor write nor create; ergo, they aren't fundamentally rational.

abstract-and-valid inference schemata

quantification

Background Claim

intensional reasoning

$\mathcal{R}$ Humans, at least neurobiologically normal ones, are fundamentally rational, where rationality is constituted by certain logico-mathematically based reasoning and decision-making in response to real-world stimuli, including stimuli given in the form of focused tests; but mere animals are not fundamentally rational, since, *contra* Darwin, their minds are fundamentally qualitatively inferior to the human mind. As to whether computing machines/robots are fundamentally rational, the answer is "No." For starters, if x can't read, write, and create, x can't be rational; computing machines/robots can neither read nor write nor create; ergo, they aren't fundamentally rational.

abstract-and-valid inference schemata

quantification

Background Claim

intensional reasoning

$\mathcal{R}$ Humans, at least neurobiologically normal ones, are fundamentally rational, where rationality is constituted by certain logico-mathematically based reasoning and decision-making in response to real-world stimuli, including stimuli given in the form of focused tests; but mere animals are not fundamentally rational, since, *contra* Darwin, their minds are fundamentally qualitatively inferior to the human mind. As to whether computing machines/robots are fundamentally rational, the answer is “No.” For starters, if [illegible] create, x can't be rational; computing machines/robots can neither read nor write nor create; ergo, they aren't fundamentally rational.

recursion

abstract-and-valid inference schemata

quantification

Background Claim

intensional reasoning

$\mathcal{R}$ Humans, at least neurobiologically normal ones, are fundamentally rational, where rationality is constituted by certain logico-mathematically based reasoning and decision-making in response to real-world stimuli, including stimuli given in the form of focused tests; but mere animals are not fundamentally rational, since, *contra* Darwin, their minds are fundamentally qualitatively inferior to the human mind. As to whether computing machines/robots are fundamentally rational, the answer is "No." For starters, if x can't read, write, and create, x can't be rational; computing machines/robots can neither read nor write nor create; ergo, they aren't fundamentally rational.

recursion

self-reference

abstract-and-valid inference schemata

quantification

intensional reasoning

recursion

self-reference

To infinity and beyond! — routinely

HS®

abstract-and-valid inference schemata

quantification

intensional reasoning

recursion

self-reference

To infinity and beyond! — routinely

abstract-and-valid inference schemata

quantification

Background Claim

intensional reasoning

HS®

$\mathcal{R}$ Humans, at least neurobiologically normal ones, are fundamentally rational, where rationality is constituted by certain logico-mathematically based reasoning and decision-making in response to real-world stimuli, including stimuli given in the form of focused tests; but mere animals are not fundamentally rational, since, *contra* Darwin, their minds are fundamentally qualitatively inferior to the human mind. As to whether computing machines/robots are fundamentally rational, the answer is "No." For starters, if x can't read, write, and create, x can't be rational; computing machines/robots can neither read nor write nor create; ergo, they aren't fundamentally rational.

recursion

self-reference

To infinity and beyond! — routinely

abstract-and-valid inference schemata

quantification

intensional reasoning

recursion

self-reference

To infinity and beyond! — routinely

abstract-and-valid inference schemata

quantification

intensional reasoning

HS®

recursion

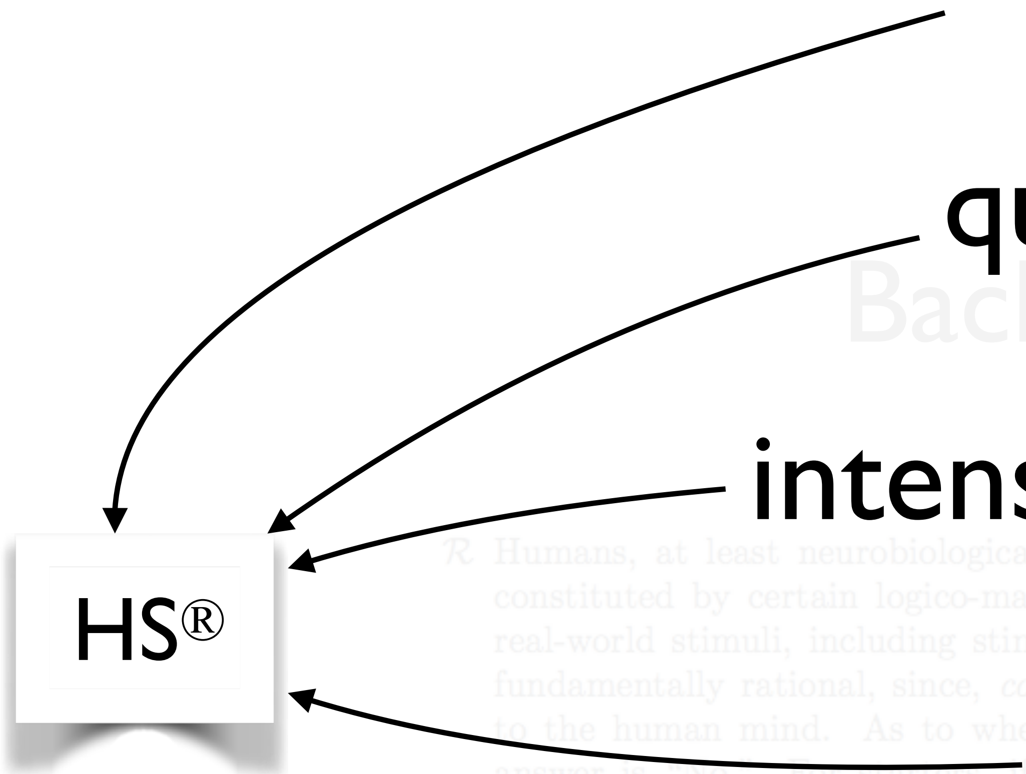
self-reference

To infinity and beyond! — routinely

abstract-and-valid inference schemata

quantification

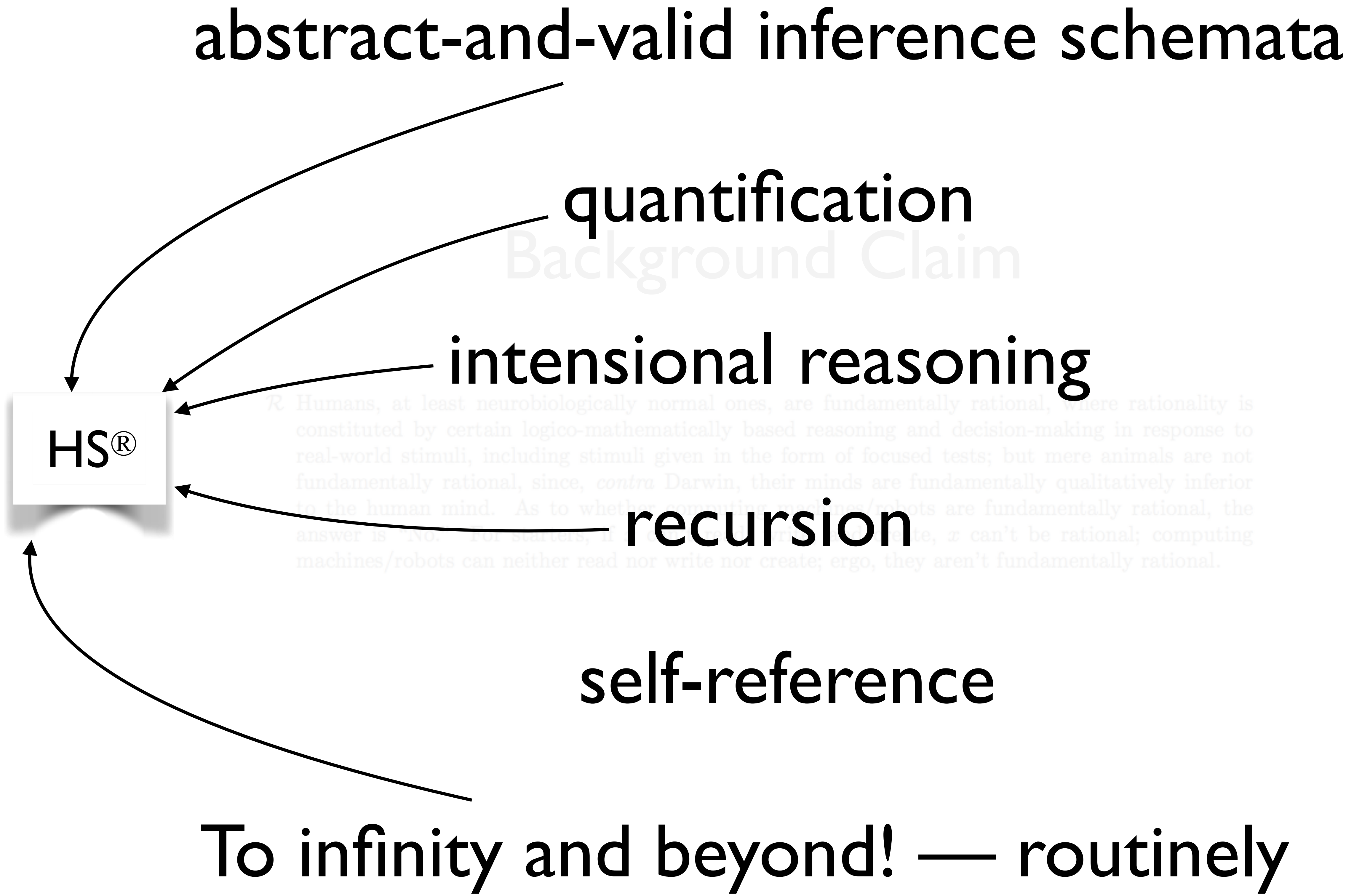
intensional reasoning

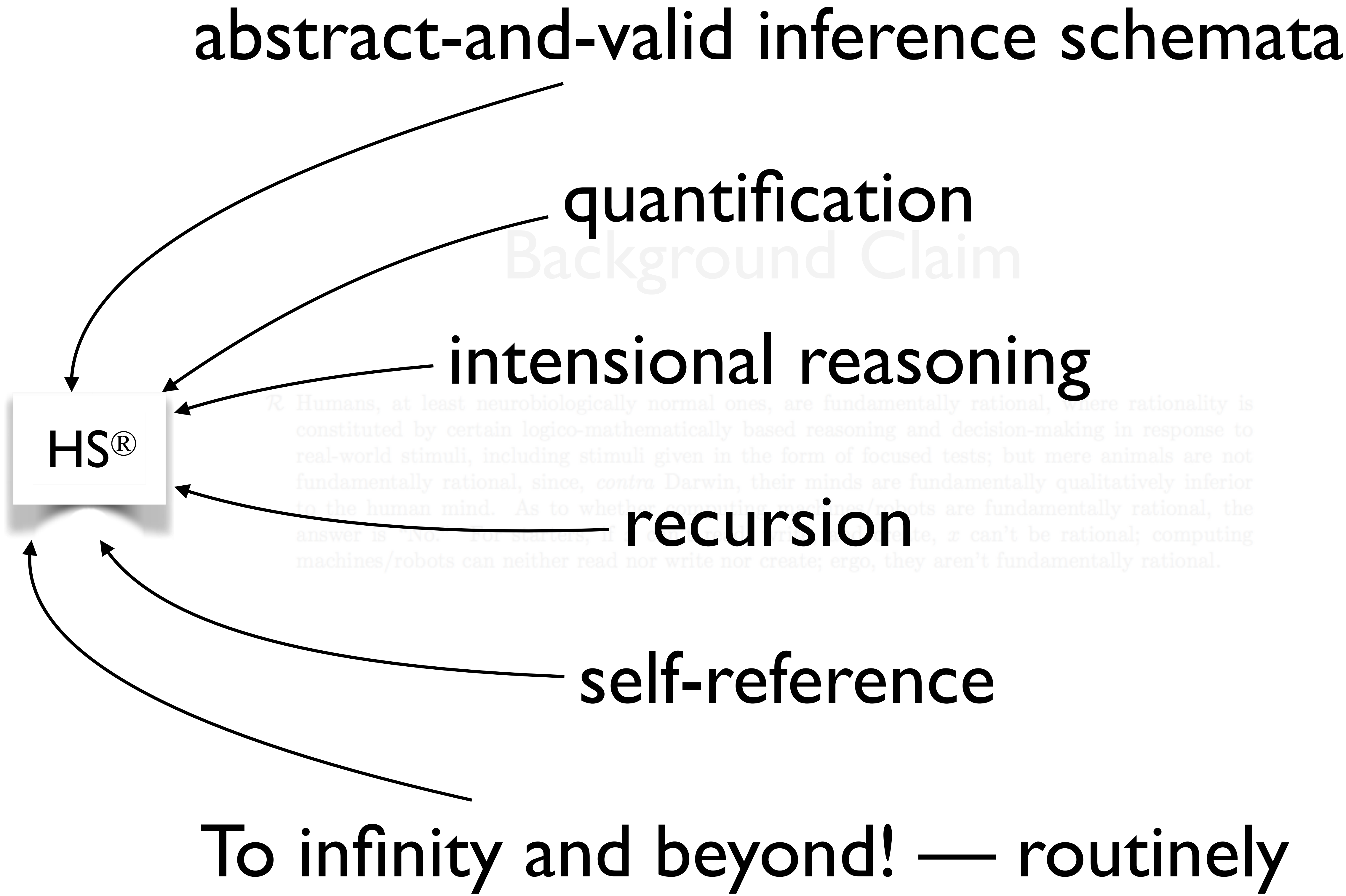


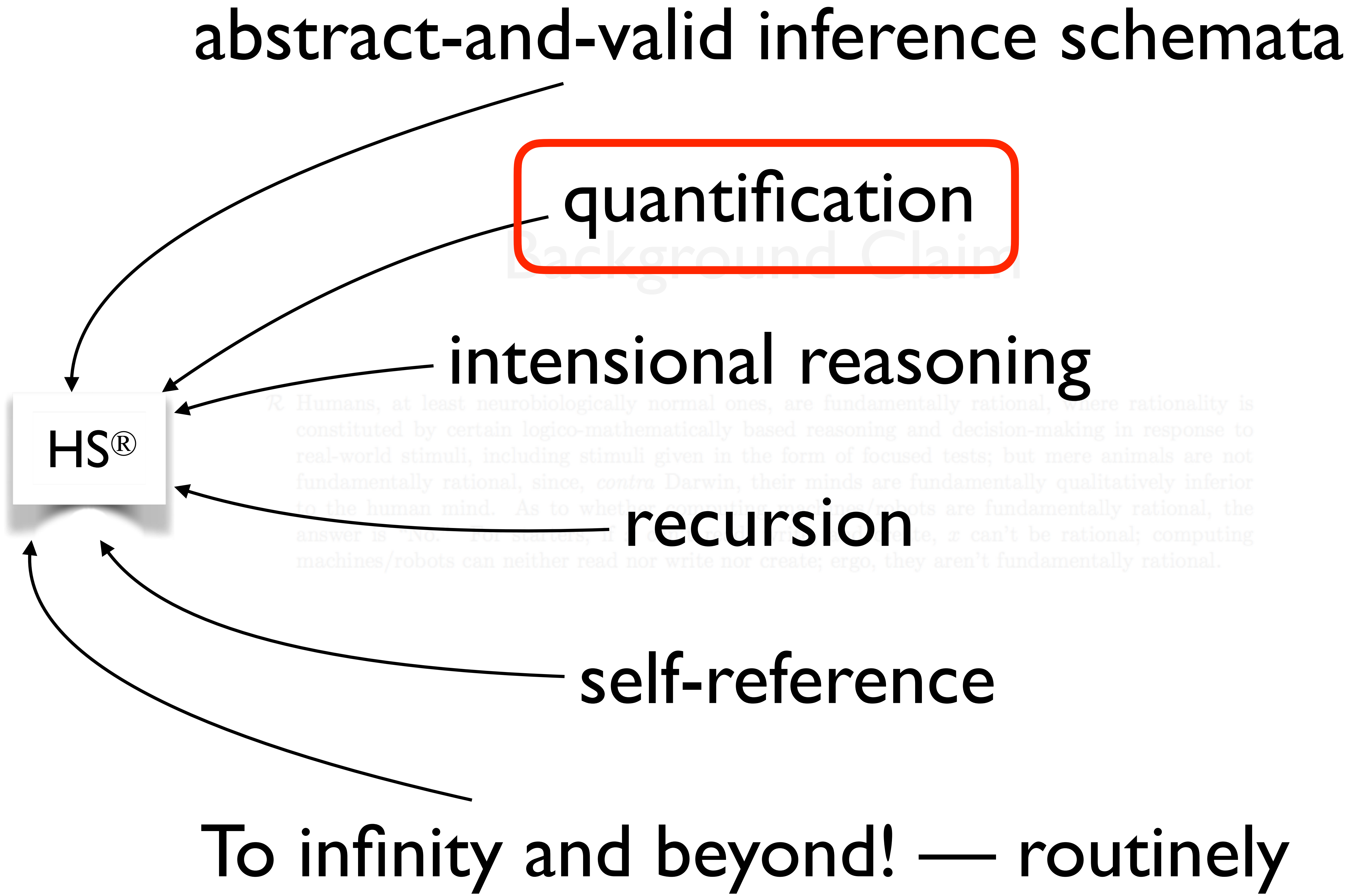
recursion

self-reference

To infinity and beyond! — routinely


abstract-and-valid inference schemata
quantification
intensional reasoning
HS®
recursion
self-reference
To infinity and beyond! — routinely


abstract-and-valid inference schemata
quantification
intensional reasoning
HS®
recursion
self-reference
To infinity and beyond! — routinely


abstract-and-valid inference schemata
quantification
intensional reasoning
HS®
recursion
self-reference
To infinity and beyond! — routinely

And now the whirlwind history …

2023

2023

$\mathcal{DCEC}^*$

Syntax

$$S ::= \text{Object} \mid \text{Agent} \mid \text{Self} \sqsubset \text{Agent} \mid \text{ActionType} \mid \text{Action} \sqsubseteq \text{Event} \mid \text{Moment} \mid \text{Boolean} \mid \text{Fluent} \mid \text{Numeric}$$

$$f ::= \begin{array}{l} \mathit{action} : \text{Agent} \times \text{ActionType} \rightarrow \text{Action} \\ \mathit{initially} : \text{Fluent} \rightarrow \text{Boolean} \\ \mathit{holds} : \text{Fluent} \times \text{Moment} \rightarrow \text{Boolean} \\ \mathit{happens} : \text{Event} \times \text{Moment} \rightarrow \text{Boolean} \\ \mathit{clipped} : \text{Moment} \times \text{Fluent} \times \text{Moment} \rightarrow \mathit{Boolean} \\ \mathit{initiates} : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Boolean} \\ \mathit{terminates} : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Boolean} \\ \mathit{prior} : \text{Moment} \times \text{Moment} \rightarrow \text{Boolean} \\ \mathit{interval} : \text{Moment} \times \text{Boolean} \\ * : \text{Agent} \rightarrow \text{Self} \\ \mathit{payoff} : \text{Agent} \times \text{ActionType} \times \text{Moment} \rightarrow \text{Numeric} \end{array}$$

$$t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n)$$

$$\phi ::= \begin{array}{l} t : \text{Boolean} \mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \\ \mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \mathbf{C}(t,\phi) \mid \mathbf{S}(a,b,t,\phi) \mid \mathbf{S}(a,t,\phi) \\ \mathbf{B}(a,t,\phi) \mid \mathbf{D}(a,t,\mathit{holds}(f,t')) \mid \mathbf{I}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t')) \\ \mathbf{O}(a,t,\phi,\mathit{happens}(\mathit{action}(a^*,\alpha),t')) \end{array}$$

Rules of Inference

$$\frac{}{\mathbf{C}(t,\mathbf{P}(a,t,\phi) \rightarrow \mathbf{K}(a,t,\phi))}\ [R_1] \qquad \frac{}{\mathbf{C}(t,\mathbf{K}(a,t,\phi) \rightarrow \mathbf{B}(a,t,\phi))}\ [R_2]$$

$$\frac{\mathbf{C}(t,\phi)\ t \leq t_1 \ldots t \leq t_n}{\mathbf{K}(a_1,t_1,\ldots \mathbf{K}(a_n,t_n,\phi)\ldots)}\ [R_3] \qquad \frac{\mathbf{K}(a,t,\phi)}{\phi}\ [R_4]$$

$$\frac{}{\mathbf{C}(t,\mathbf{K}(a,t_1,\phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{K}(a,t_2,\phi_1) \rightarrow \mathbf{K}(a,t_3,\phi_2)}\ [R_5]$$

$$\frac{}{\mathbf{C}(t,\mathbf{B}(a,t_1,\phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{B}(a,t_2,\phi_1) \rightarrow \mathbf{B}(a,t_3,\phi_2)}\ [R_6]$$

$$\frac{}{\mathbf{C}(t,\mathbf{C}(t_1,\phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{C}(t_2,\phi_1) \rightarrow \mathbf{C}(t_3,\phi_2)}\ [R_7]$$

$$\frac{}{\mathbf{C}(t,\forall x.\ \phi \rightarrow \phi[x \mapsto t])}\ [R_8] \qquad \frac{}{\mathbf{C}(t,\phi_1 \leftrightarrow \phi_2 \rightarrow \neg\phi_2 \rightarrow \neg\phi_1)}\ [R_9]$$

$$\frac{}{\mathbf{C}(t,[\phi_1 \wedge \ldots \wedge \phi_n \rightarrow \phi] \rightarrow [\phi_1 \rightarrow \ldots \rightarrow \phi_n \rightarrow \psi])}\ [R_{10}]$$

$$\frac{\mathbf{B}(a,t,\phi)\quad \phi \rightarrow \psi}{\mathbf{B}(a,t,\psi)}\ [R_{11a}] \qquad \frac{\mathbf{B}(a,t,\phi)\quad \mathbf{B}(a,t,\psi)}{\mathbf{B}(a,t,\psi \wedge \phi)}\ [R_{11b}]$$

$$\frac{\mathbf{S}(s,h,t,\phi)}{\mathbf{B}(h,t,\mathbf{B}(s,t,\phi))}\ [R_{12}]$$

$$\frac{\mathbf{I}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t'))}{\mathbf{P}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t))}\ [R_{13}]$$

$$\frac{\mathbf{B}(a,t,\phi)\quad \mathbf{B}(a,t,\mathbf{O}(a^*,t,\phi,\mathit{happens}(\mathit{action}(a^*,\alpha),t')))\quad \mathbf{O}(a,t,\phi,\mathit{happens}(\mathit{action}(a^*,\alpha),t'))}{\mathbf{K}(a,t,\mathbf{I}(a^*,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t')))}\ [R_{14}]$$

$$\frac{\phi \leftrightarrow \psi}{\mathbf{O}(a,t,\phi,\gamma) \leftrightarrow \mathbf{O}(a,t,\psi,\gamma)}\ [R_{15}]$$

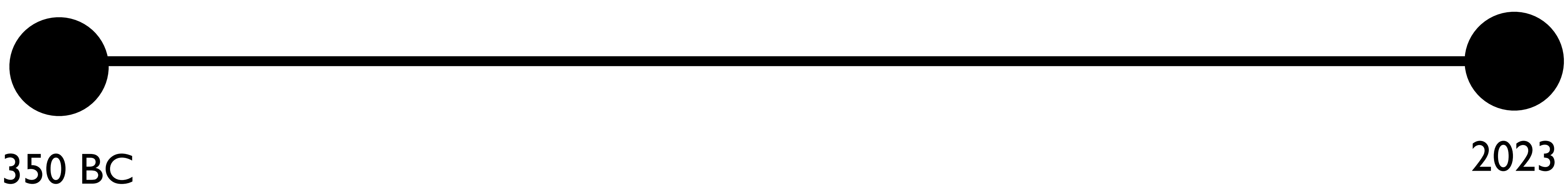
2023

Intro to (Formal) Logic (& AI) @ RPI

2023


2023


2023


350 BC
2023

350 BC

2023

Euclid

Euclidean "Magic"

Theorem: There are infinitely many primes.

Proof: We take an indirect route. Let $\Pi = p_1 = 2, p_2 = 3, p_3 = 5, \ldots, p_k$ be a finite, exhaustive consecutive sequence of prime numbers. Next, let $\mathbf{M}_\Pi$ be $p_1 \times p_2 \times \cdots \times p_k$, and set $\mathbf{M}'_\Pi$ to $\mathbf{M}_\Pi + 1$. Either $\mathbf{M}'_\Pi$ is prime, or not; we thus have two (exhaustive) cases to consider.

C1 Suppose $\mathbf{M}'_\Pi$ is prime. In this case we immediately have a prime number beyond any in Π — contradiction!

C2 Suppose on the other hand that $\mathbf{M}'_\Pi$ is *not* prime. Then some prime p divides $\mathbf{M}'_\Pi$. (Why?) Now, p itself is either in Π, or not; we hence have two sub-cases. Supposing that p is in Π entails that p divides $\mathbf{M}_\Pi$. But we are operating under the supposition that p divides $\mathbf{M}'_\Pi$ as well. This implies that p divides 1, which is absurd (a contradiction). Hence the prime p is outside Π.

Hence for *any* such list Π, there is a prime outside the list. That is, there are infinitely many primes. **QED**

Euclidean "Magic"

Theorem: There are infinitely many primes.

Proof: We take an indirect route. Let $\Pi = p_1 = 2, p_2 = 3, p_3 = 5, \ldots, p_k$ be a finite, exhaustive consecutive sequence of prime numbers. Next, let $\mathbf{M}_\Pi$ be $p_1 \times p_2 \times \cdots \times p_k$, and set $\mathbf{M}'_\Pi$ to $\mathbf{M}_\Pi + 1$. Either $\mathbf{M}'_\Pi$ is prime, or not; we thus have two (exhaustive) cases to consider.

C1 Suppose $\mathbf{M}'_\Pi$ is prime. In this case we immediately have a prime number beyond any in Π — contradiction!

C2 Suppose on the other hand that $\mathbf{M}'_\Pi$ is *not* prime. Then some prime p divides $\mathbf{M}'_\Pi$. (Why?) Now, p itself is either in Π, or not; we hence have two sub-cases. Supposing that p is in Π entails that p divides $\mathbf{M}_\Pi$. But we are operating under the supposition that p divides $\mathbf{M}'_\Pi$ as well. This implies that p divides 1, which is absurd (a contradiction). Hence the prime p is outside Π.

Hence for *any* such list Π, there is a prime outside the list. That is, there are infinitely many primes. **QED**

Euclidean "Magic"

Theorem: There are infinitely many primes.

Proof: We take an indirect route. Let $\Pi = p_1 = 2, p_2 = 3, p_3 = 5, \ldots, p_k$ be a finite, exhaustive consecutive sequence of prime numbers. Next, let $\mathbf{M}_\Pi$ be $p_1 \times p_2 \times \cdots \times p_k$, and set $\mathbf{M}'_\Pi$ to $\mathbf{M}_\Pi + 1$. Either $\mathbf{M}'_\Pi$ is prime, or not; we thus have two (exhaustive) cases to consider.

C1 Suppose $\mathbf{M}'_\Pi$ is prime. In this case we immediately have a prime number beyond any in Π — contradiction!

C2 Suppose on the other hand that $\mathbf{M}'_\Pi$ is *not* prime. Then some prime p divides $\mathbf{M}'_\Pi$. (Why?) Now, p itself is either in Π, or not; we hence have two sub-cases. Supposing that p is in Π entails that p divides $\mathbf{M}_\Pi$. But we are operating under the supposition that p divides $\mathbf{M}'_\Pi$ as well. This implies that p divides 1, which is absurd (a contradiction). Hence the prime p is outside Π.

Hence for *any* such list Π, there is a prime outside the list. That is, there are infinitely many primes. **QED**

Euclidean "Magic"

Theorem: There are infinitely many primes.

Proof: We take an indirect route. Let $\Pi = p_1 = 2, p_2 = 3, p_3 = 5, \ldots, p_k$ be a finite, exhaustive consecutive sequence of prime numbers. Next, let $\mathbf{M}_\Pi$ be $p_1 \times p_2 \times \cdots \times p_k$, and set $\mathbf{M}'_\Pi$ to $\mathbf{M}_\Pi + 1$. Either $\mathbf{M}'_\Pi$ is prime, or not; we thus have two (exhaustive) cases to consider.

C1 Suppose $\mathbf{M}'_\Pi$ is prime. In this case we immediately have a prime number beyond any in Π — contradiction!

C2 Suppose on the other hand that $\mathbf{M}'_\Pi$ is *not* prime. Then some prime p divides $\mathbf{M}'_\Pi$. (Why?) Now, p itself is either in Π, or not; we hence have two sub-cases. Supposing that p is in Π entails that p divides $\mathbf{M}_\Pi$. But we are operating under the supposition that p divides $\mathbf{M}'_\Pi$ as well. This implies that p divides 1, which is absurd (a contradiction). Hence the prime p is outside Π.

Hence for *any* such list Π, there is a prime outside the list. That is, there are infinitely many primes. **QED**

Euclidean "Magic"

Theorem: There are infinitely many primes.

Proof: We take an indirect route. Let $\Pi = p_1 = 2, p_2 = 3, p_3 = 5, \ldots, p_k$ be a finite, exhaustive consecutive sequence of prime numbers. Next, let $\mathbf{M}_\Pi$ be $p_1 \times p_2 \times \cdots \times p_k$, and set $\mathbf{M}'_\Pi$ to $\mathbf{M}_\Pi + 1$. Either $\mathbf{M}'_\Pi$ is prime, or not; we thus have two (exhaustive) cases to consider.

C1 Suppose $\mathbf{M}'_\Pi$ is prime. In this case we immediately have a prime number beyond any in Π — contradiction!

C2 Suppose on the other hand that $\mathbf{M}'_\Pi$ is *not* prime. Then some prime p divides $\mathbf{M}'_\Pi$. (Why?) Now, p itself is either in Π, or not; we hence have two sub-cases. Supposing that p is in Π entails that p divides $\mathbf{M}_\Pi$. But we are operating under the supposition that p divides $\mathbf{M}'_\Pi$ as well. This implies that p divides 1, which is absurd (a contradiction). Hence the prime p is outside Π.

Hence for *any* such list Π, there is a prime outside the list. That is, there are infinitely many primes. **QED**

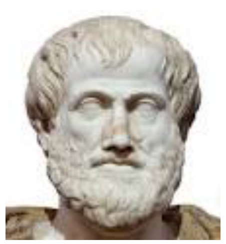
The Fundamental Theorem of Arithmetic

350 BC

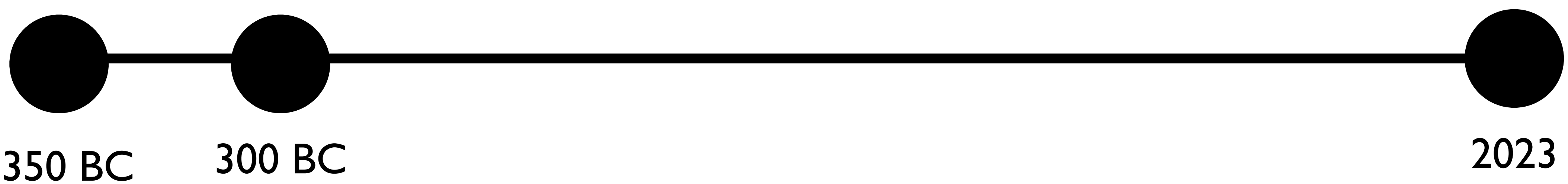
2023

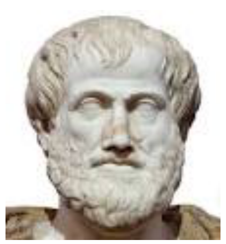
Euclid

350 BC 300 BC 2023

Euclid



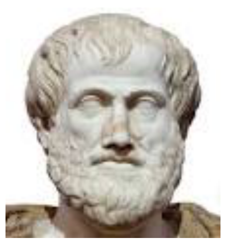


I don’t believe in magic! Why exactly is that so convincing? What exactly is he doing?!?

Euclid

350 BC 300 BC 2023

Euclid



Organon

I don’t believe in magic! Why exactly is that so convincing? What exactly is he doing?!?

He's using syllogisms!

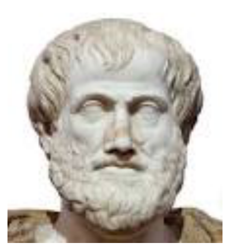
E.g.,

All As are Bs.
All Bs are Cs.

All As are Cs.

350 BC | 300 BC | 2023

Euclid



Organon

I don't believe in magic! Why exactly is that so convincing? What exactly is he doing?!?

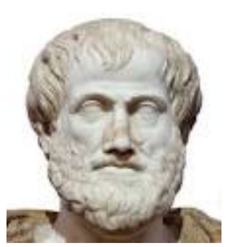
He’s using syllogisms!

E.g.,

All As are Bs.
All Bs are Cs.

All As are Cs.

Euclid



Organon

I don’t believe in magic! Why exactly is that so convincing? What exactly is he doing?!?

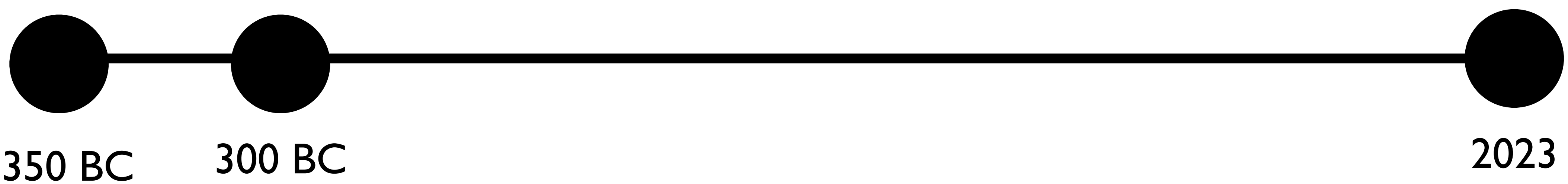
Balderdash!

He's using syllogisms!

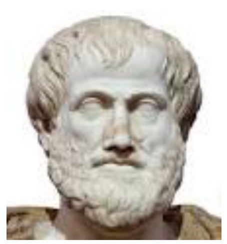
E.g.,

All As are Bs.
All Bs are Cs.

All As are Cs.



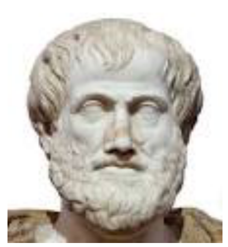
Euclid



Organon

I don't believe in magic! Why exactly is that so convincing? What exactly is he doing?!?

Balderdash!



Euclid

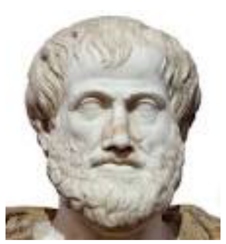
Organon

I don't believe in magic! Why exactly is that so convincing? What exactly is he doing?!?

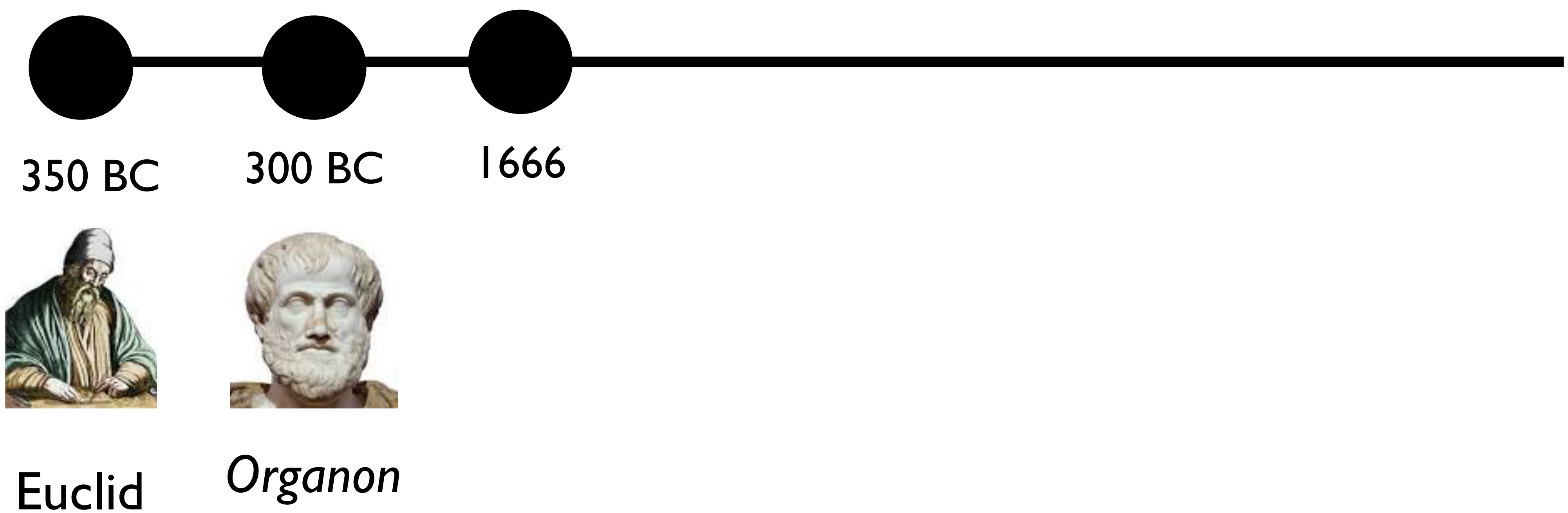
350 BC

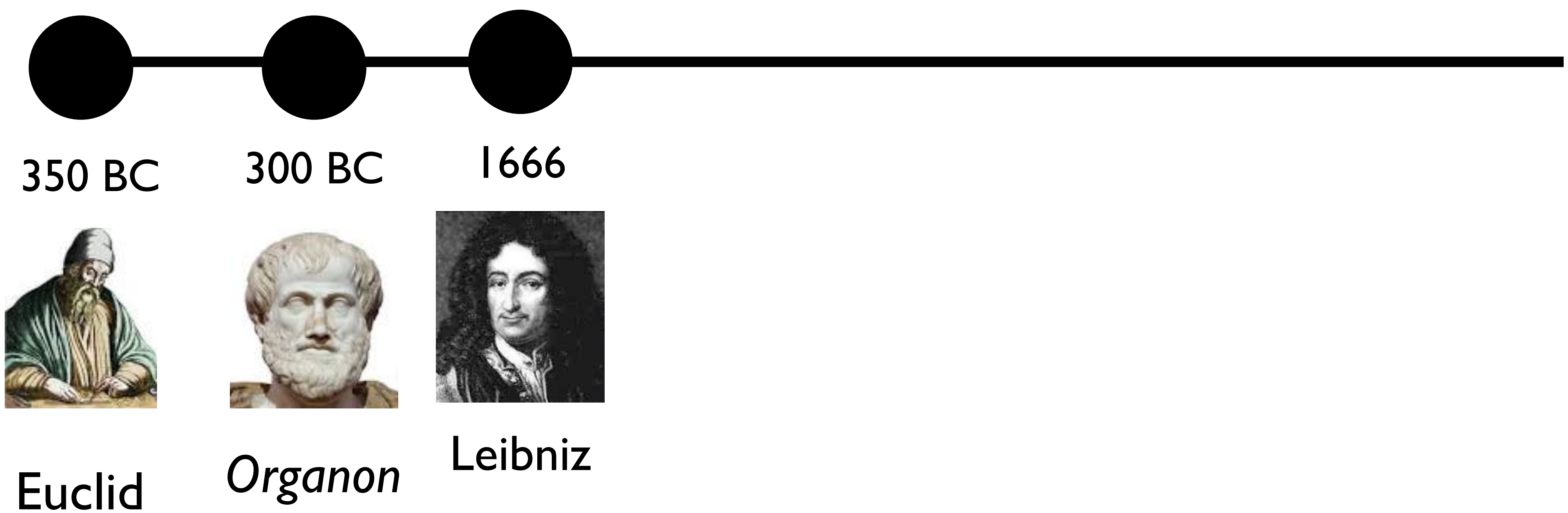
Euclid

300 BC



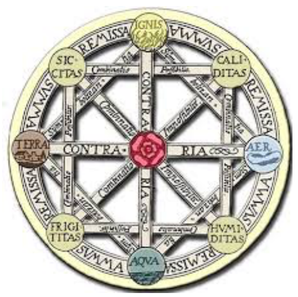
Organon


350 BC
300 BC
1666
Euclid
Organon


350 BC
300 BC
1666
Euclid
Organon
Leibniz


350 BC
300 BC
1666
Euclid
Organon
Leibniz
∫

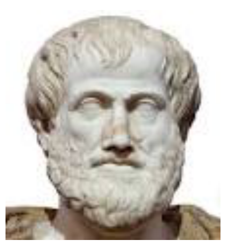
"Universal Computational Logic"



350 BC

Euclid

300 BC



Organon

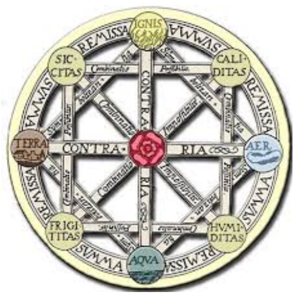
1666

Leibniz

$\int$

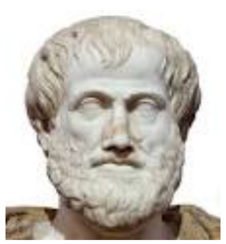
Intro to (Formal) Logic (& AI) @ RPI

"Universal Computational Logic"



350 BC | 300 BC | 1666 | 1854

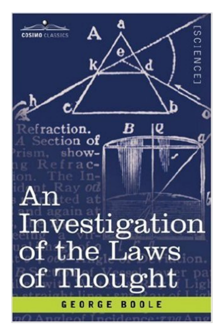
Euclid



Organon

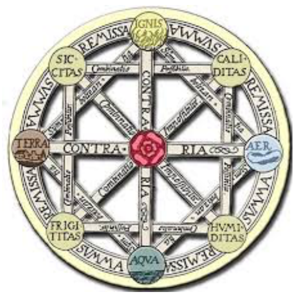
Leibniz

$\int$



Intro to (Formal) Logic (& AI) @ RPI

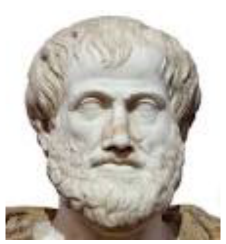
"Universal Computational Logic"



350 BC

Euclid

300 BC



Organon

1666

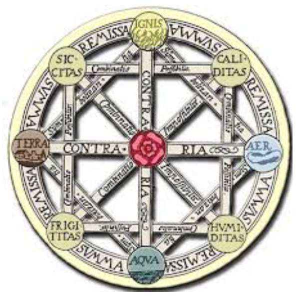
Leibniz

$\int$

1854

Intro to (Formal) Logic (& AI) @ RPI

"Universal Computational Logic"

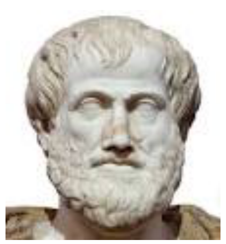


Logic Theorist
(birth of modern logicist AI)

350 BC

Euclid

300 BC



Organon

1666

Leibniz

$\int$

1854

1956

Simon

"Astonishing" Logic Theorist Proof @ Dawn of AI

“Astonishing” Logic Theorist Proof @ Dawn of AI

1	$(\phi \vee \phi) \rightarrow \phi$	axiom
2	$(\neg\phi \vee \neg\phi) \rightarrow \neg\phi$	substitution
3	$(\phi \rightarrow \neg\phi) \rightarrow \neg\phi$	a “replacement rule”
4	$(A \rightarrow \neg A) \rightarrow \neg A$	substitution

“Astonishing” Logic Theorist Proof @ Dawn of AI

1	$(\phi \vee \phi) \rightarrow \phi$	axiom
2	$(\neg\phi \vee \neg\phi) \rightarrow \neg\phi$	substitution
3	$(\phi \rightarrow \neg\phi) \rightarrow \neg\phi$	a “replacement rule”
4	$(A \rightarrow \neg A) \rightarrow \neg A$	substitution

At dawn of AI: 10 seconds.

“Astonishing” Logic Theorist Proof @ Dawn of AI

1	$(\phi \vee \phi) \rightarrow \phi$	axiom
2	$(\neg\phi \vee \neg\phi) \rightarrow \neg\phi$	substitution
3	$(\phi \rightarrow \neg\phi) \rightarrow \neg\phi$	a “replacement rule”
4	$(A \rightarrow \neg A) \rightarrow \neg A$	substitution

At dawn of AI: 10 seconds.

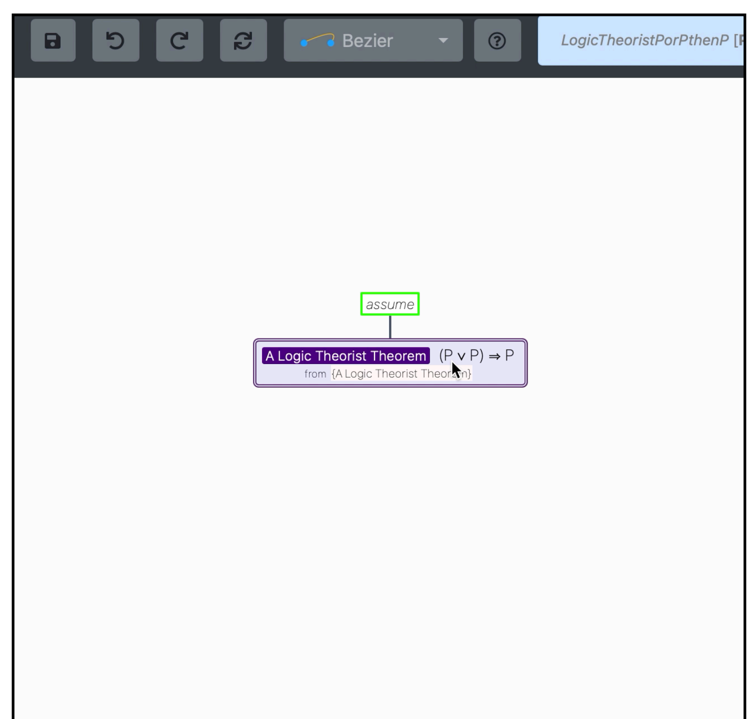
AI of today: vanishingly small amount of time (in eg HS®).

“Astonishing” Logic Theorist Proof @ Dawn of AI

1	$(\phi \vee \phi) \rightarrow \phi$	axiom
2	$(\neg\phi \vee \neg\phi) \rightarrow \neg\phi$	substitution
3	$(\phi \rightarrow \neg\phi) \rightarrow \neg\phi$	a “replacement rule”
4	$(A \rightarrow \neg A) \rightarrow \neg A$	substitution

At dawn of AI: 10 seconds.

AI of today: vanishingly small amount of time (in eg HS®).

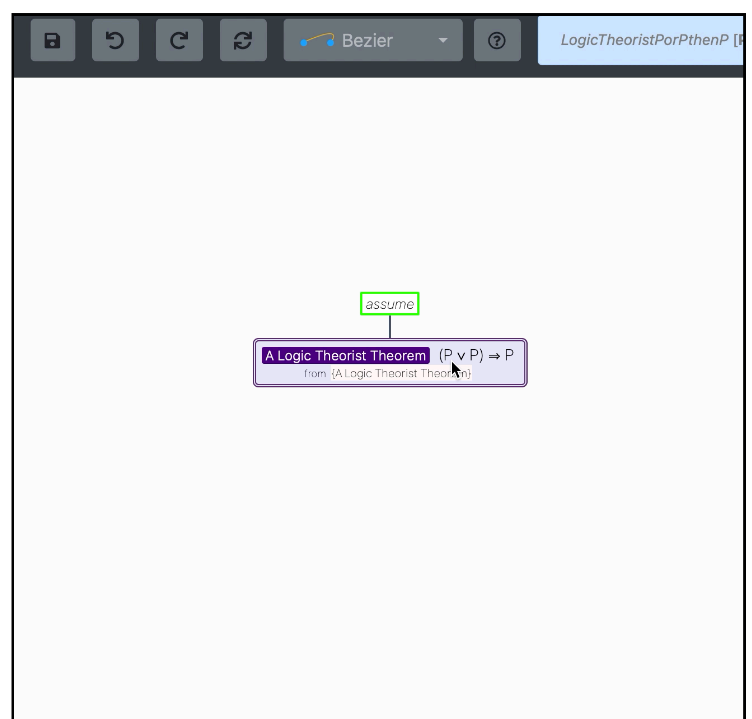


"Astonishing" Logic Theorist Proof @ Dawn of AI

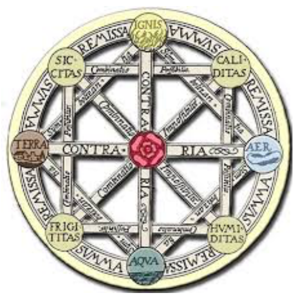
1	$(\phi \vee \phi) \rightarrow \phi$	axiom
2	$(\neg\phi \vee \neg\phi) \rightarrow \neg\phi$	substitution
3	$(\phi \rightarrow \neg\phi) \rightarrow \neg\phi$	a "replacement rule"
4	$(A \rightarrow \neg A) \rightarrow \neg A$	substitution

At dawn of AI: 10 seconds.

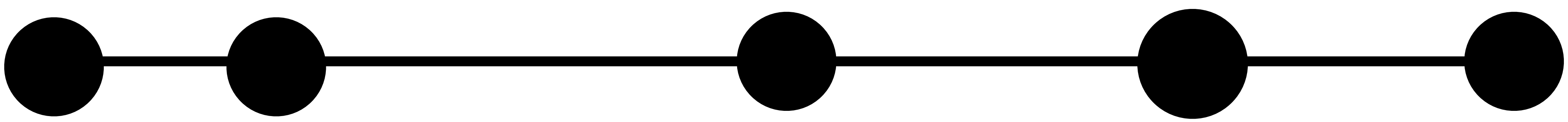
AI of today: vanishingly small amount of time (in eg HS®).



"Universal Computational Logic"



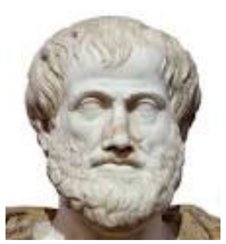
Logic Theorist
(birth of modern logicist AI)



350 BC

Euclid

300 BC

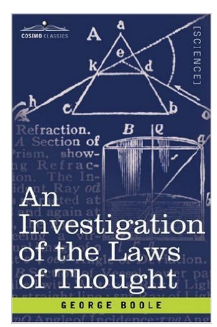


Organon

Leibniz

$\int$

1854



1956

Simon

2023

Intro to (Formal) Logic (& AI) @ RPI

"Universal Computational Logic"

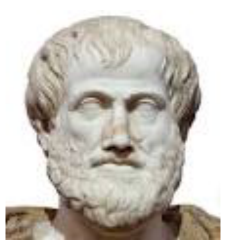
Logic Theorist
(birth of modern logicist AI)

350 BC — 300 BC — 1666 — 1854 — 1956 — 2023

350 BC

Euclid

300 BC



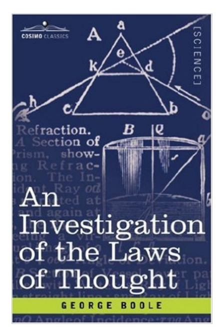
Organon

1666

Leibniz

$\int$

1854



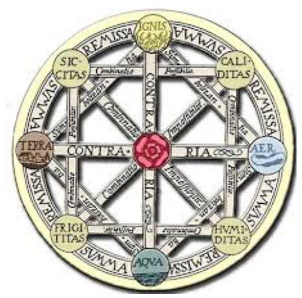
1956

Simon

2023

Intro to (Formal) Logic (& AI) @ RPI

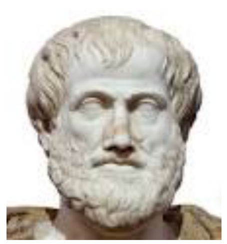
“Universal Computational Logic”



Logic Theorist
(birth of modern logicist AI)

350 BC	300 BC	1666	1854	1956	2023

Euclid



Organon

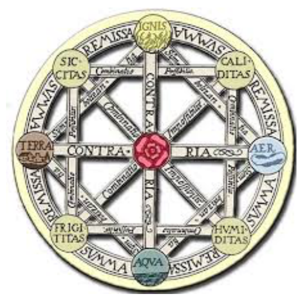
Leibniz

$\int$

Simon

Intro to (Formal) Logic (& AI) @ RPI

“Universal Computational Logic”

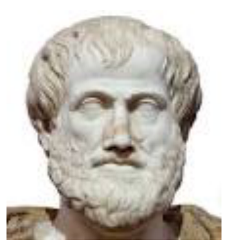


Logic Theorist

(birth of modern logicist AI)

350 BC	300 BC	1666	1854	1956	2023	2024

Euclid



Organon

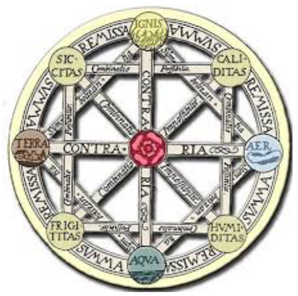
Leibniz

$\int$

Simon

Intro to (Formal) Logic (& AI) @ RPI

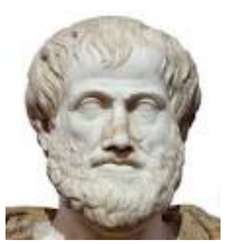
"Universal Computational Logic"



Logic Theorist
(birth of modern logicist AI)

350 BC	300 BC	1666	1854	1956	2023	2024

Euclid



Organon

Leibniz

$\int$

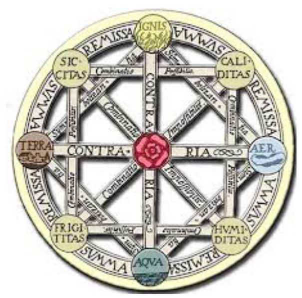
Simon

The Singularity?

Intro to (Formal) Logic (& AI) @ RPI

Entscheidungsproblem

“Universal Computational Logic”

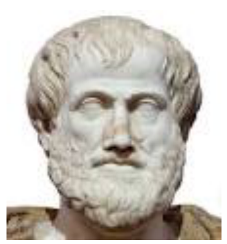


Logic Theorist
(birth of modern logicist AI)

350 BC

Euclid

300 BC



Organon

1666

Leibniz

$\int$

1854

1956

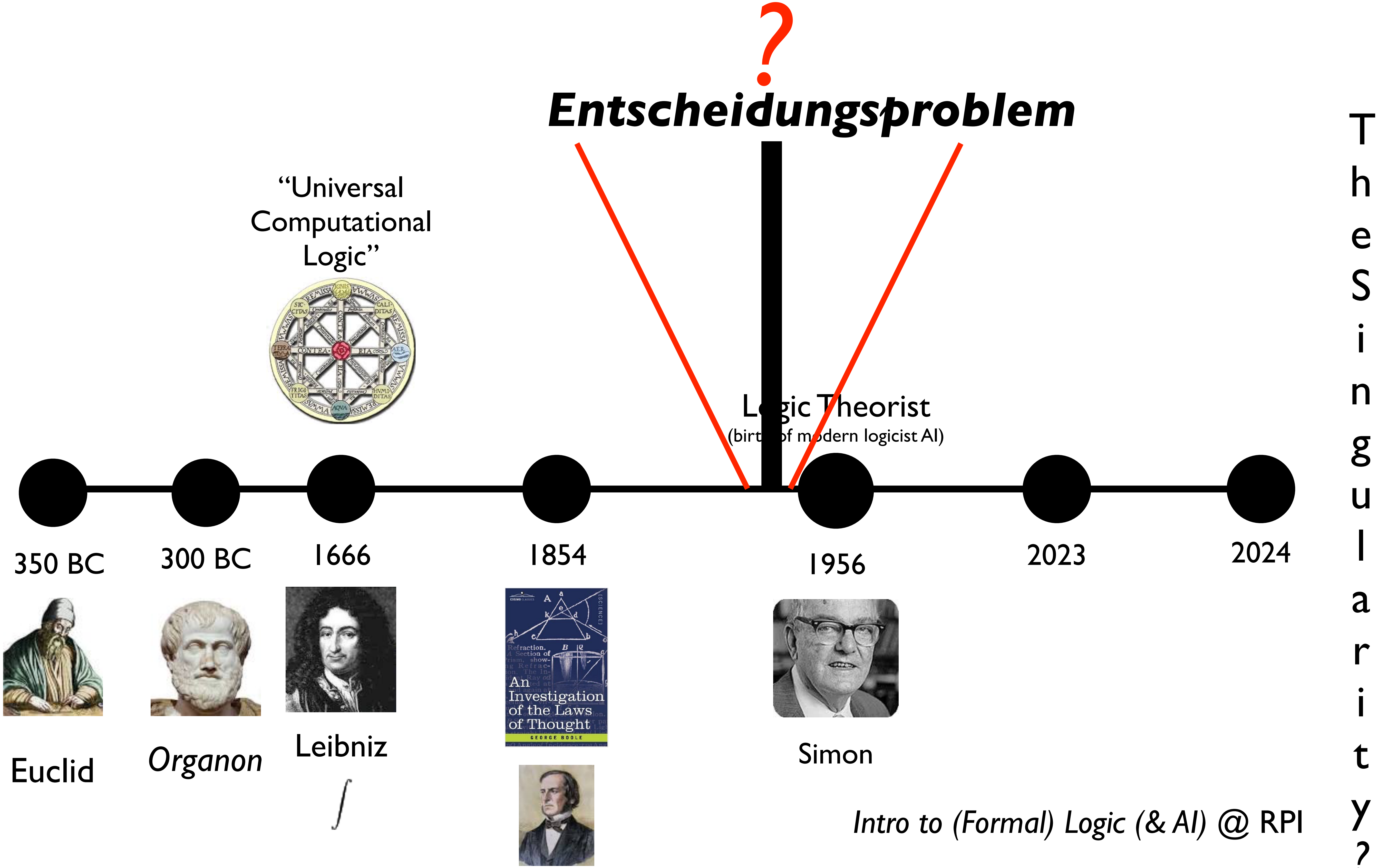
Simon

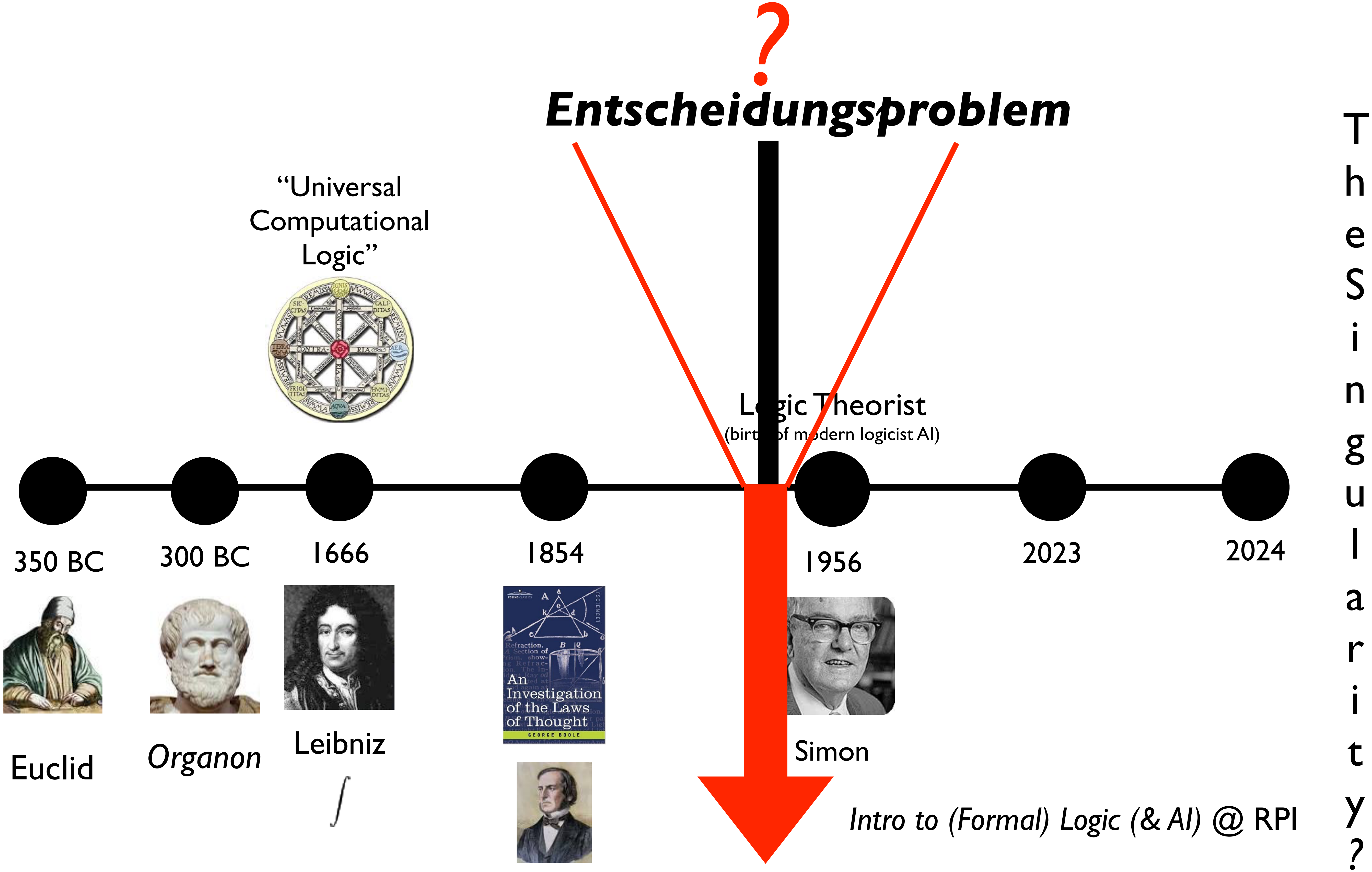
2023

2024

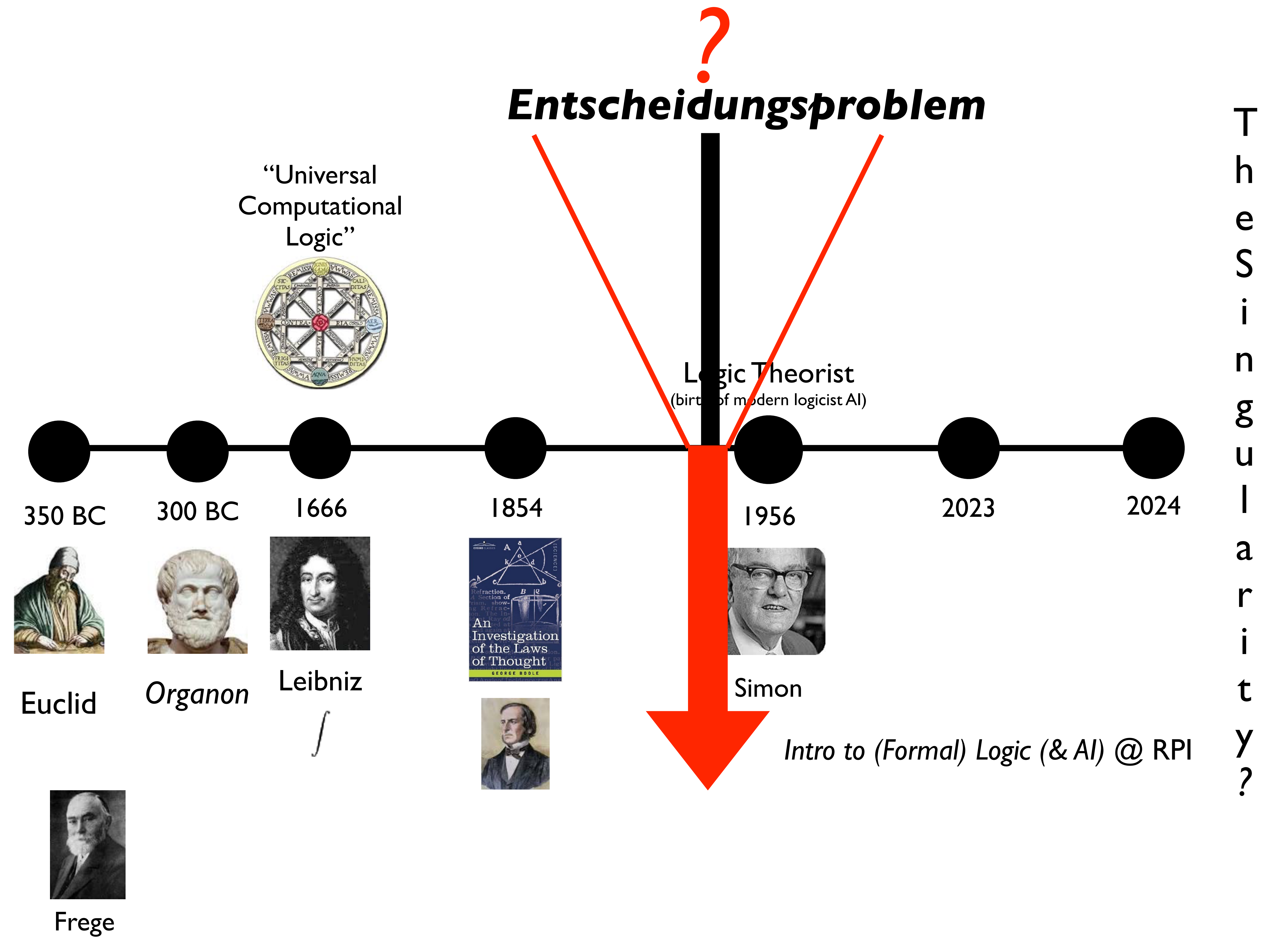
The Singularity?

Intro to (Formal) Logic (& AI) @ RPI

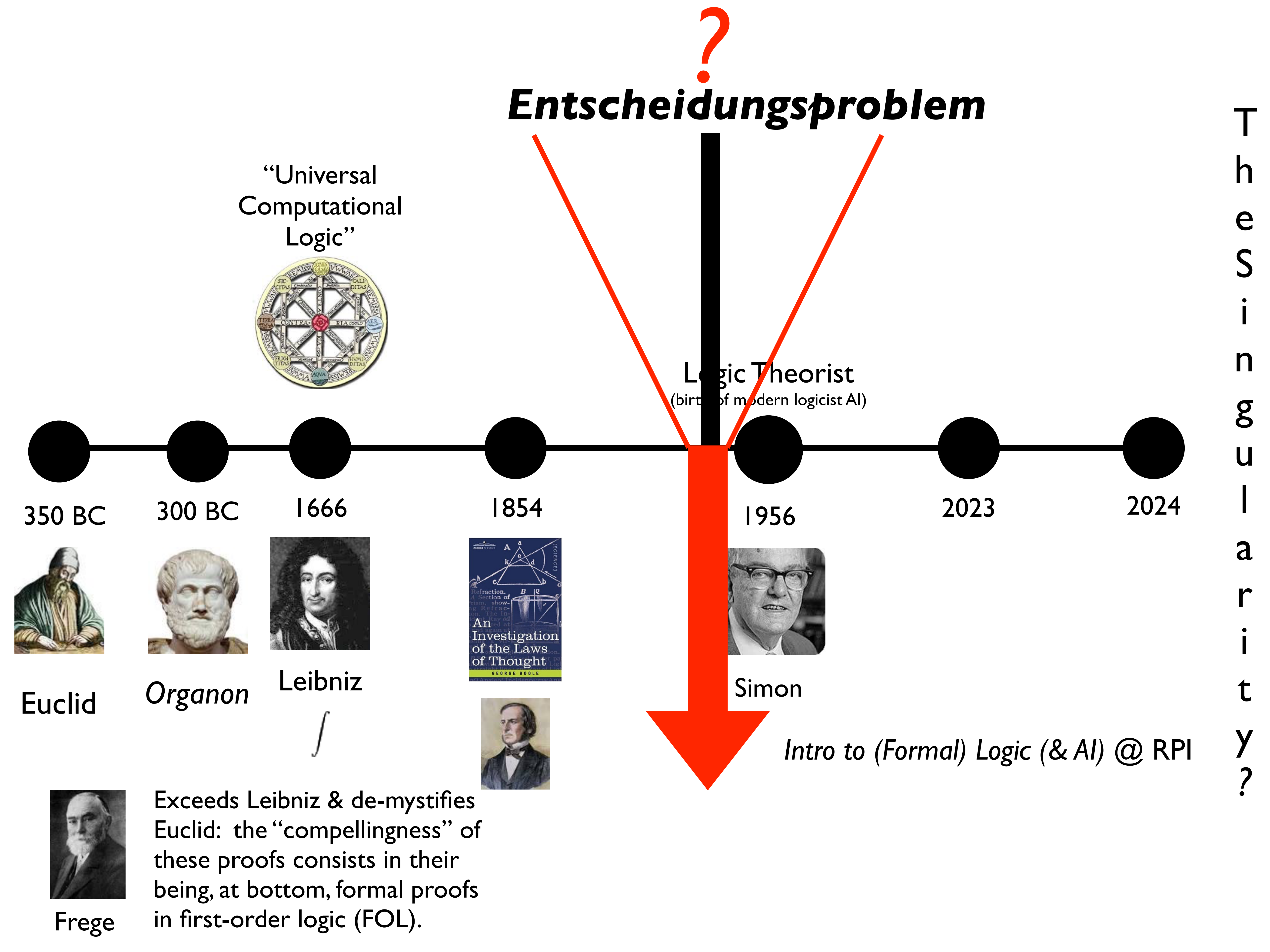

?
Entscheidungsproblem
"Universal Computational Logic"
Logic Theorist
(birth of modern logicist AI)
350 BC
300 BC
1666
1854
1956
2023
2024
Euclid
Organon
Leibniz
∫
An Investigation of the Laws of Thought
George Boole
Simon
Intro to (Formal) Logic (& AI) @ RPI
The Singularity?

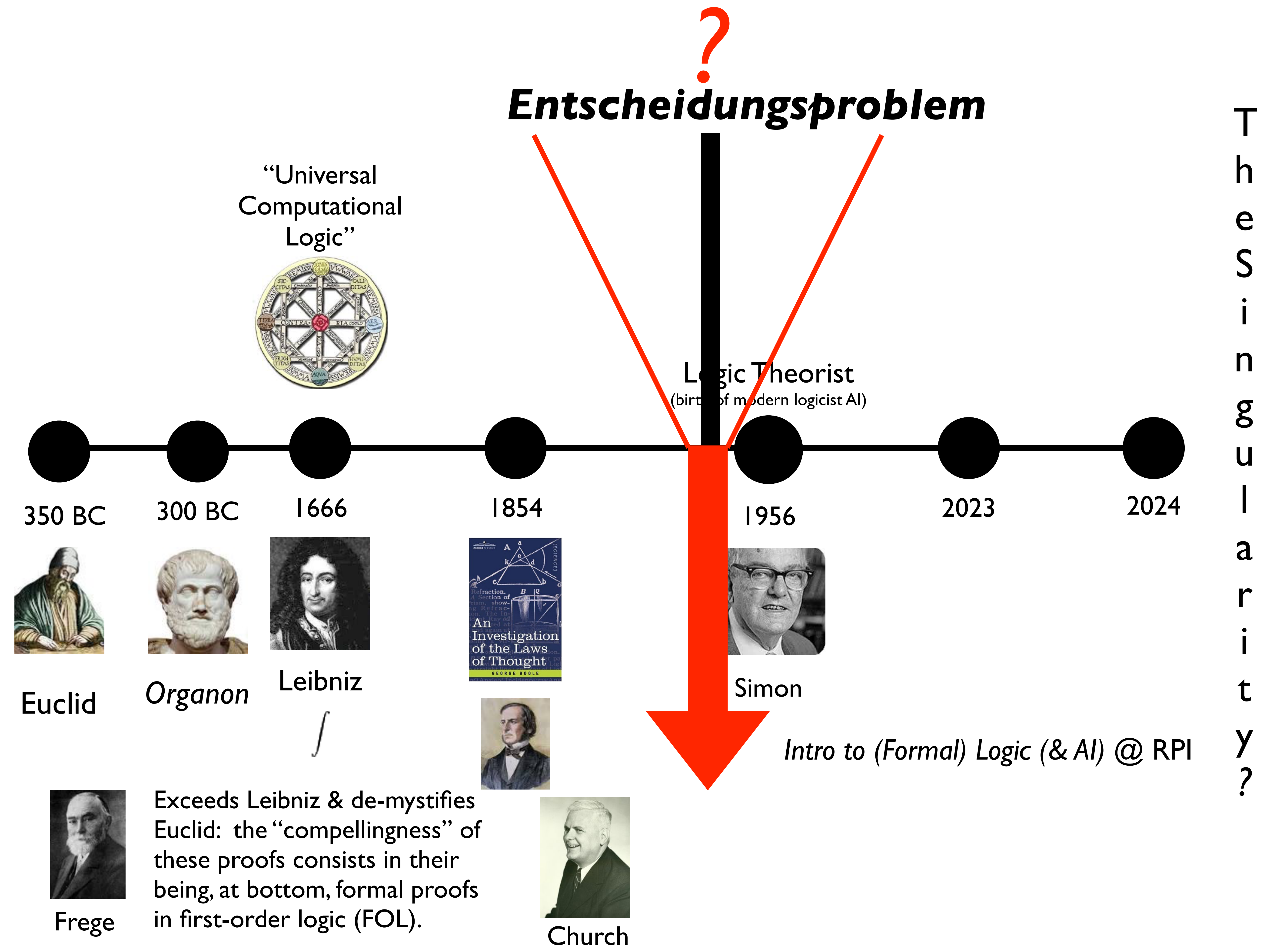


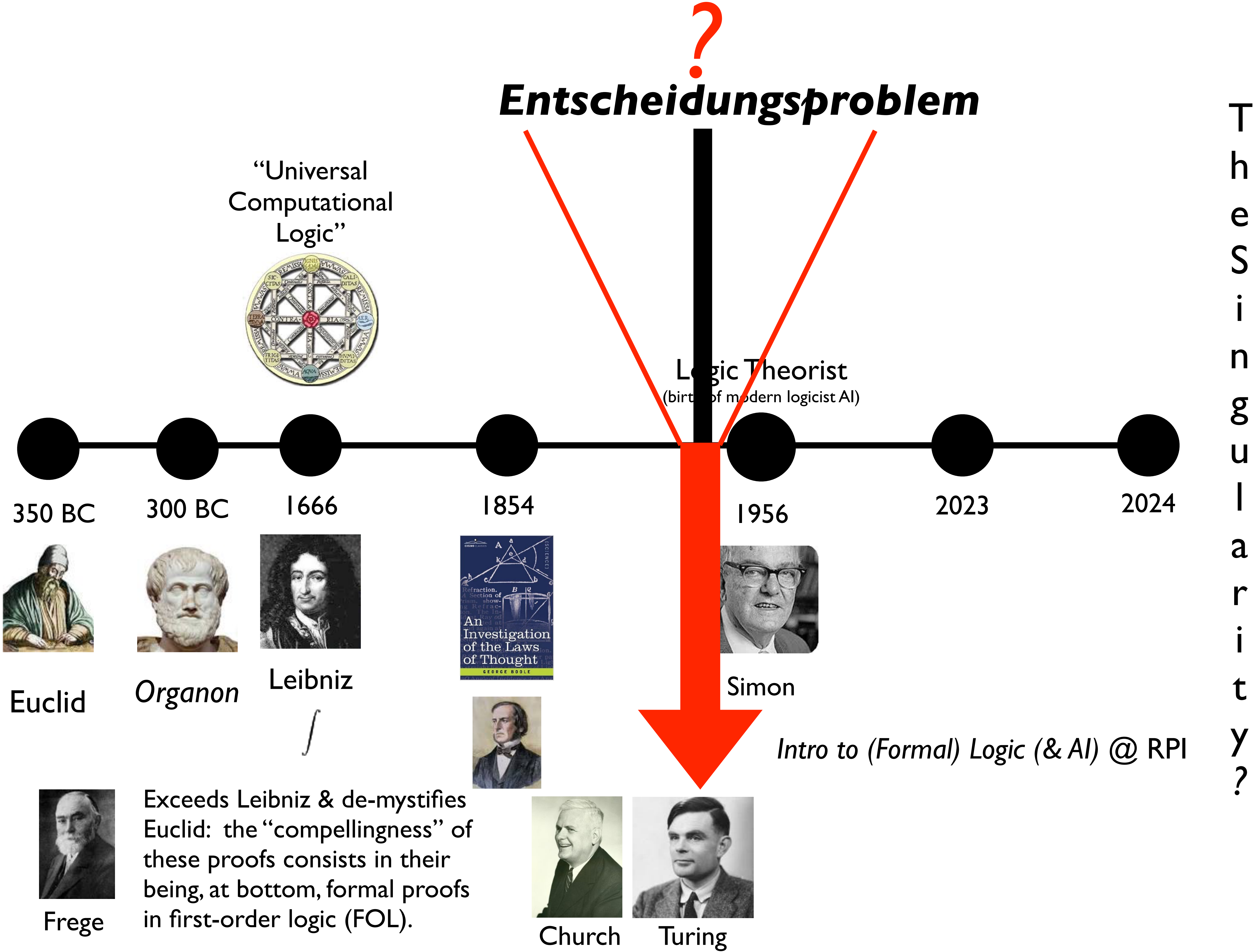
?
Entscheidungsproblem
“Universal Computational Logic”
Logic Theorist
(birth of modern logicist AI)
350 BC
300 BC
1666
1854
1956
2023
2024
Euclid
Organon
Leibniz
∫
An Investigation of the Laws of Thought
GEORGE BOOLE
Simon
Intro to (Formal) Logic (& AI) @ RPI
The Singularity?

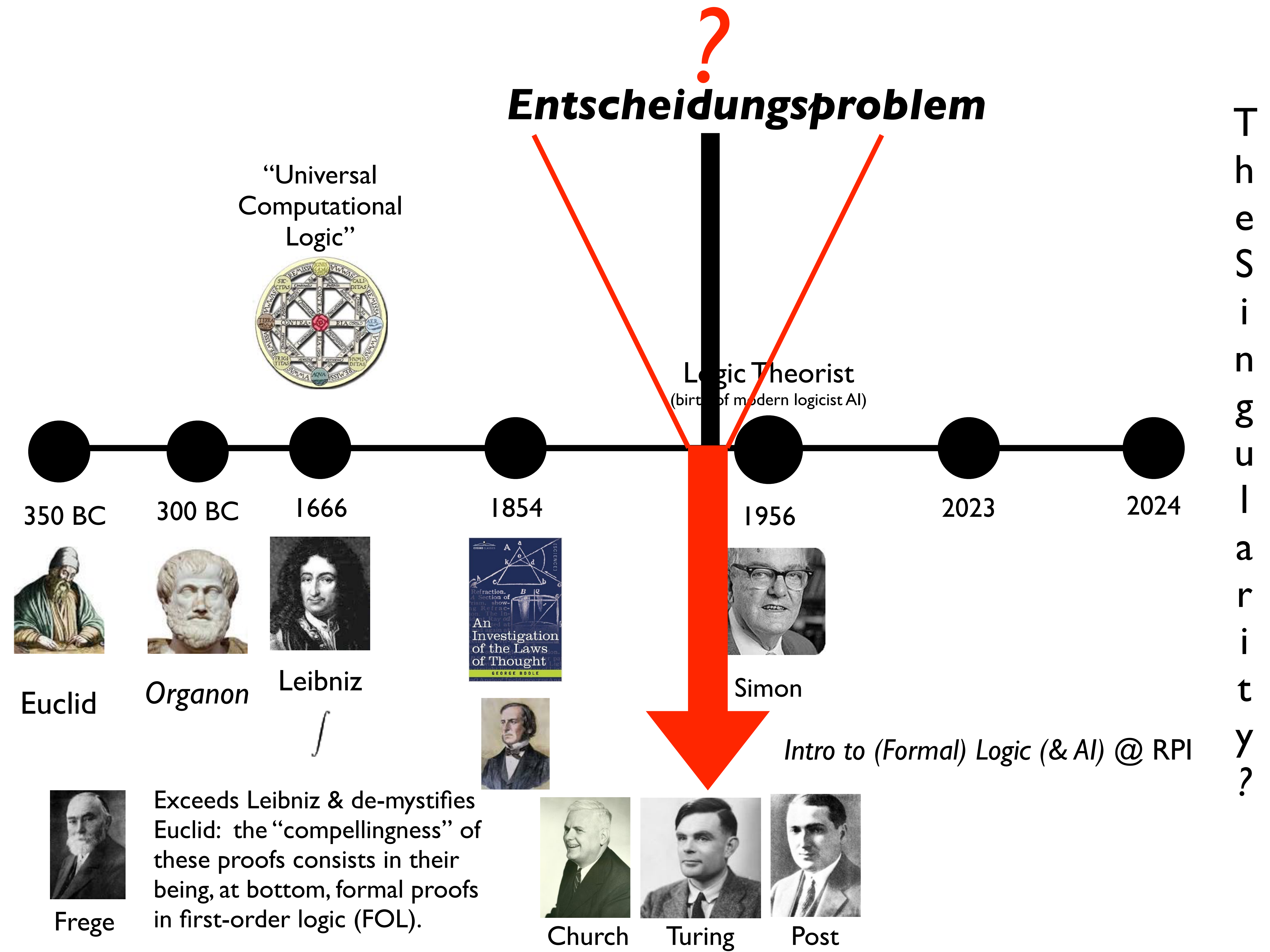


Intro to (Formal) Logic (& AI) @ RPI


?
Entscheidungsproblem
"Universal Computational Logic"
Logic Theorist
(birth of modern logicist AI)
350 BC
300 BC
1666
1854
1956
2023
2024
Euclid
Organon
Leibniz
An Investigation of the Laws of Thought
GEORGE BOOLE
Simon
Intro to (Formal) Logic (& AI) @ RPI
The Singularity?
Frege
Exceeds Leibniz & de-mystifies Euclid: the "compellingness" of these proofs consists in their being, at bottom, formal proofs in first-order logic (FOL).


?
Entscheidungsproblem
"Universal Computational Logic"
Logic Theorist
(birth of modern logicist AI)
350 BC
300 BC
1666
1854
1956
2023
2024
Euclid
Organon
Leibniz
∫
An Investigation of the Laws of Thought
GEORGE BOOLE
Simon
Intro to (Formal) Logic (& AI) @ RPI
Frege
Exceeds Leibniz & de-mystifies Euclid: the "compellingness" of these proofs consists in their being, at bottom, formal proofs in first-order logic (FOL).
Church
The Singularity?


?
Entscheidungsproblem
T h e S i n g u l a r i t y ?
"Universal Computational Logic"
Logic Theorist
(birth of modern logicist AI)
350 BC
300 BC
1666
1854
1956
2023
2024
Euclid
Organon
Leibniz
∫
An Investigation of the Laws of Thought
GEORGE BOOLE
Simon
Intro to (Formal) Logic (& AI) @ RPI
Frege
Exceeds Leibniz & de-mystifies Euclid: the "compellingness" of these proofs consists in their being, at bottom, formal proofs in first-order logic (FOL).
Church
Turing


?
Entscheidungsproblem
"Universal Computational Logic"
Logic Theorist
(birth of modern logicist AI)
350 BC
300 BC
1666
1854
1956
2023
2024
Euclid
Organon
Leibniz
∫
An Investigation of the Laws of Thought
GEORGE BOOLE
Simon
Intro to (Formal) Logic (& AI) @ RPI
Frege
Exceeds Leibniz & de-mystifies Euclid: the "compellingness" of these proofs consists in their being, at bottom, formal proofs in first-order logic (FOL).
Church
Turing
Post
The Singularity?

?

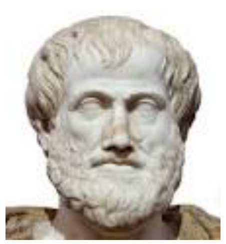
Entscheidungsproblem

“Universal Computational Logic”

Logic Theorist
(birth of modern logicist AI)

350 BC — 300 BC — 1666 — 1854 — 1956 — 2023 — 2024

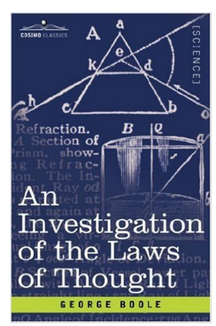
Euclid



Organon

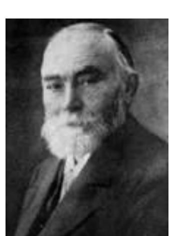
Leibniz

$\int$



Simon

Intro to (Formal) Logic (& AI) @ RPI



Frege

Exceeds Leibniz & de-mystifies Euclid: the “compellingness” of these proofs consists in their being, at bottom, formal proofs in first-order logic (FOL).

Church

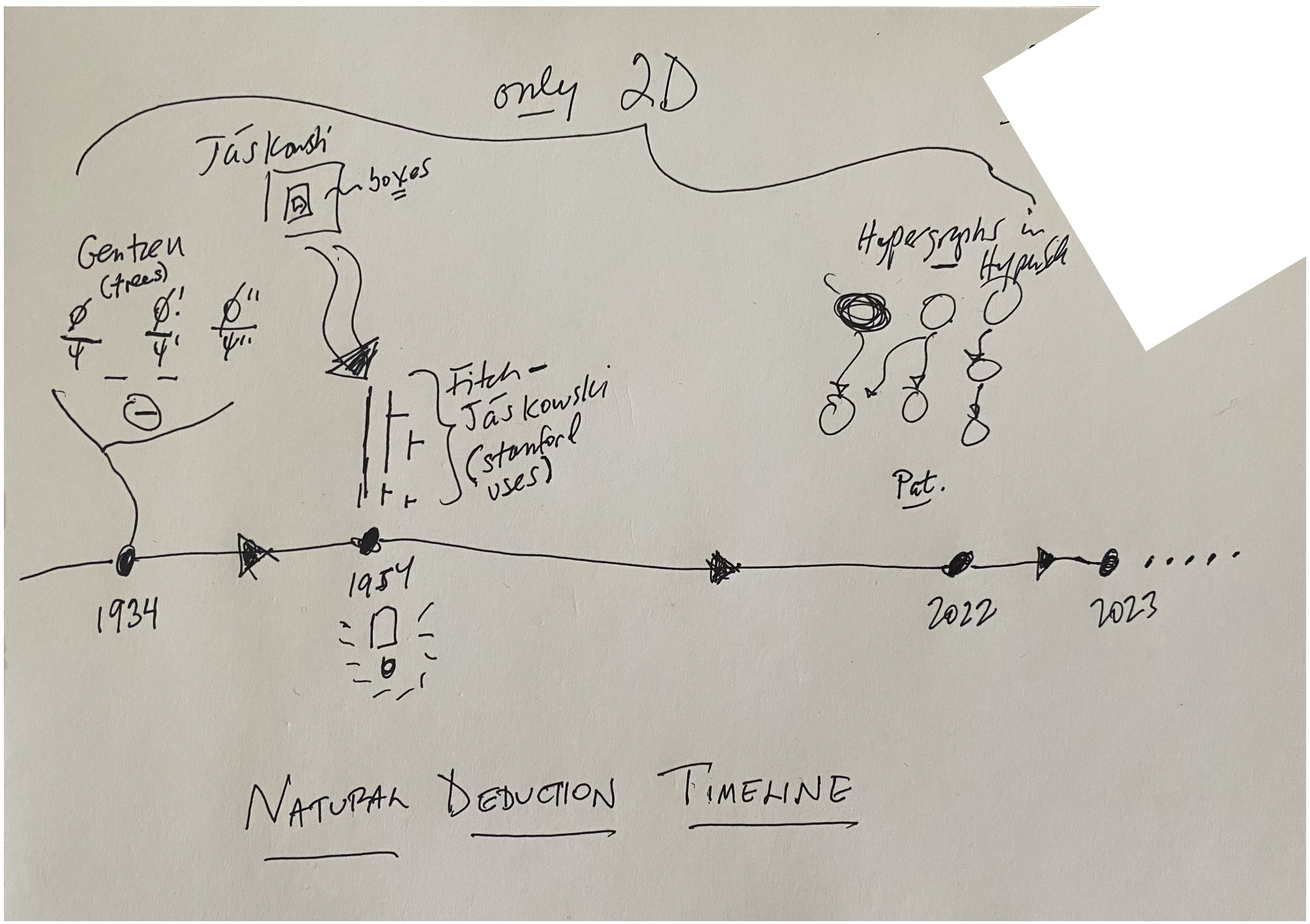
Turing

Post

Here’s what a computer is, and given that, sorry, the *Entscheidungsproblem* can’t be solved by such a machine!

The Singularity?

A Sub-History, for Later



First, the Theoremhood Decision Problem (THEOREM$_{PC}$) for the Propositional Calculus

First, the Theoremhood Decision Problem ($\text{THEOREM}_{\text{PC}}$) for the Propositional Calculus

First, the Theoremhood Decision Problem (THEOREM_{PC}) for the Propositional Calculus

First, the Theoremhood Decision Problem ($THEOREM_{PC}$) for the Propositional Calculus

First, the Theoremhood Decision Problem ($\text{THEOREM}_{\text{PC}}$) for the Propositional Calculus

First, the Theoremhood Decision Problem (THEOREM_{PC}) for the Propositional Calculus

First, the Theoremhood Decision Problem (THEOREM$_{PC}$) for the Propositional Calculus

Hard!! — for apparently no polynomial-time algorithm for this!

First, the Theoremhood Decision Problem ($\text{THEOREM}_{\text{PC}}$) for the Propositional Calculus

Hard!! — for apparently no polynomial-time algorithm for this!

First, the Theoremhood Decision Problem (THEOREM_{PC}) for the Propositional Calculus

$$((K \to A) \land \neg A) \to \neg K$$

Hard!! — for apparently no polynomial-time algorithm for this!

First, the Theoremhood Decision Problem ($\text{THEOREM}_{\text{PC}}$) for the Propositional Calculus

$$((K \rightarrow A) \wedge \neg A) \rightarrow \neg K$$

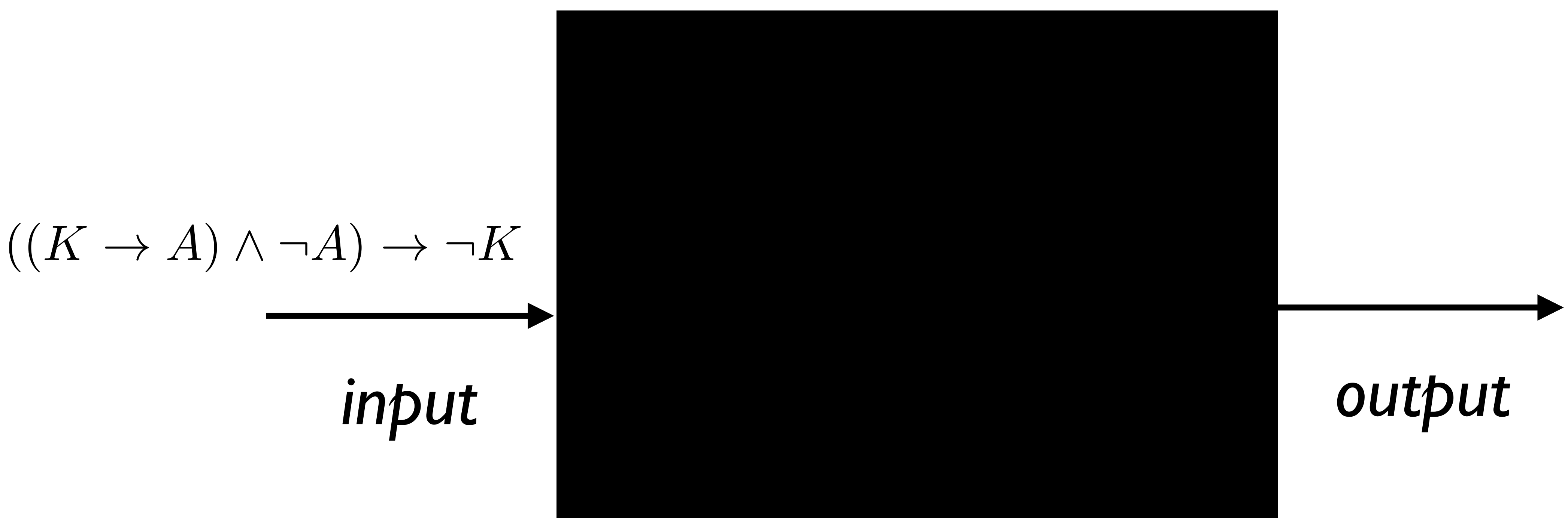
Hard!! — for apparently no polynomial-time algorithm for this!

First, the Theoremhood Decision Problem ($\text{THEOREM}_{\text{PC}}$) for the Propositional Calculus

$$((K \to A) \wedge \neg A) \to \neg K$$

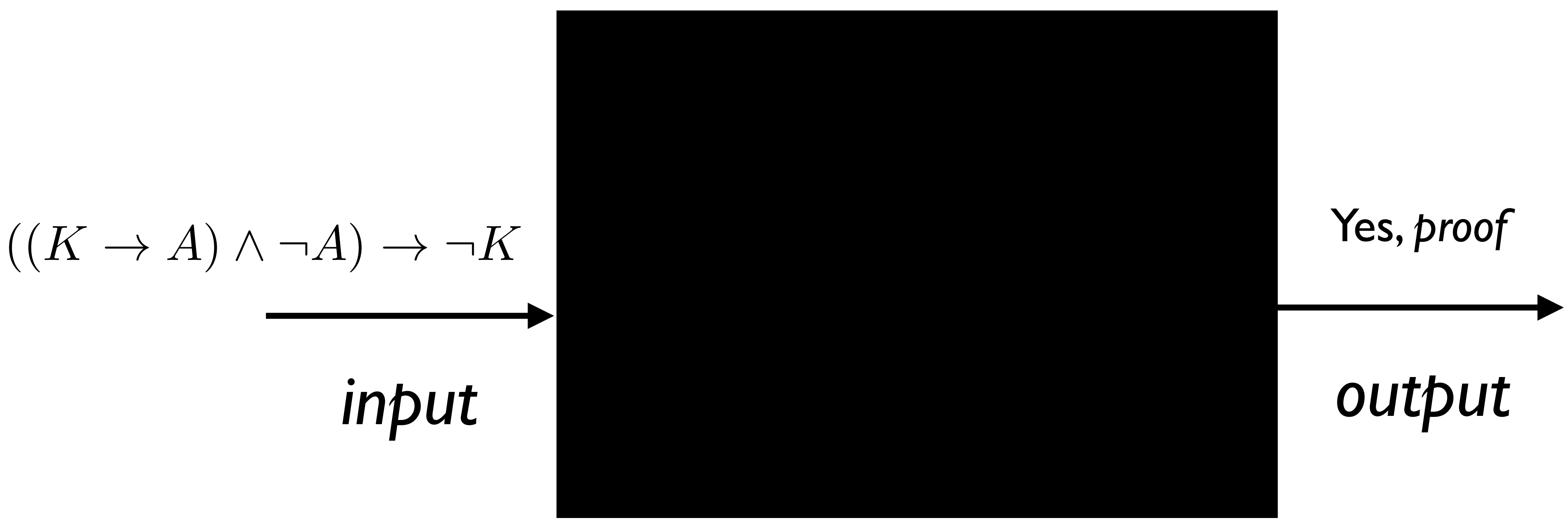
Hard!! — for apparently no polynomial-time algorithm for this!

First, the Theoremhood Decision Problem (THEOREM_{PC}) for the Propositional Calculus



Hard!! — for apparently no polynomial-time algorithm for this!

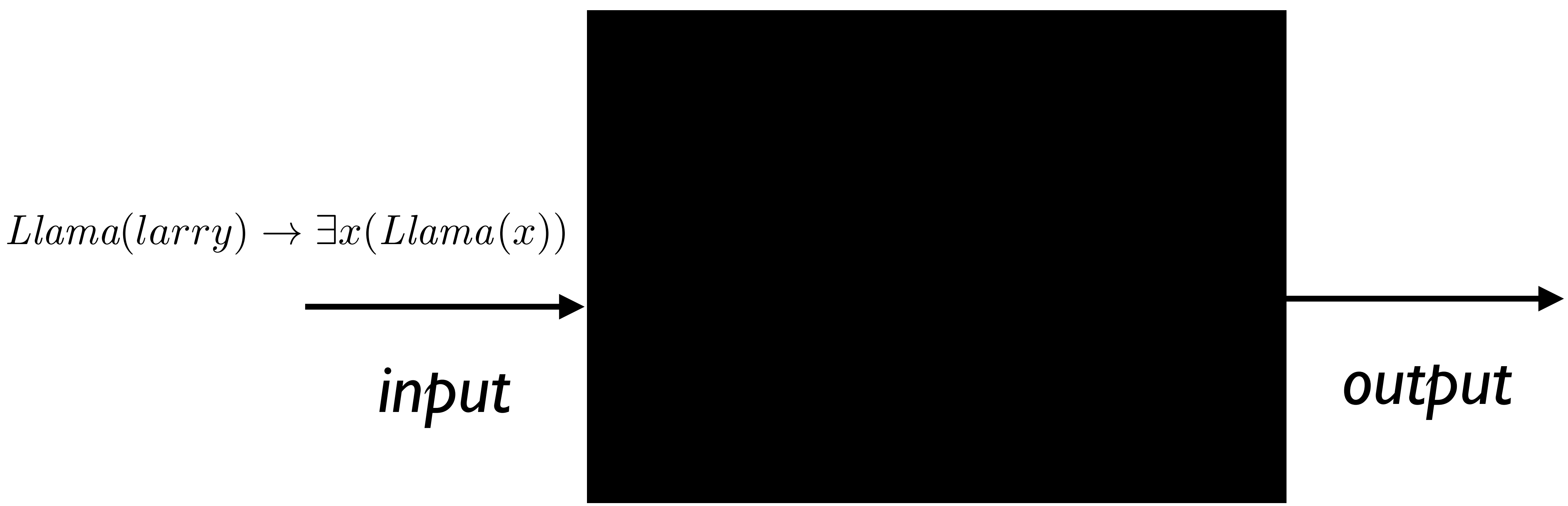
First, the Theoremhood Decision Problem (THEOREM$_{PC}$) for the Propositional Calculus



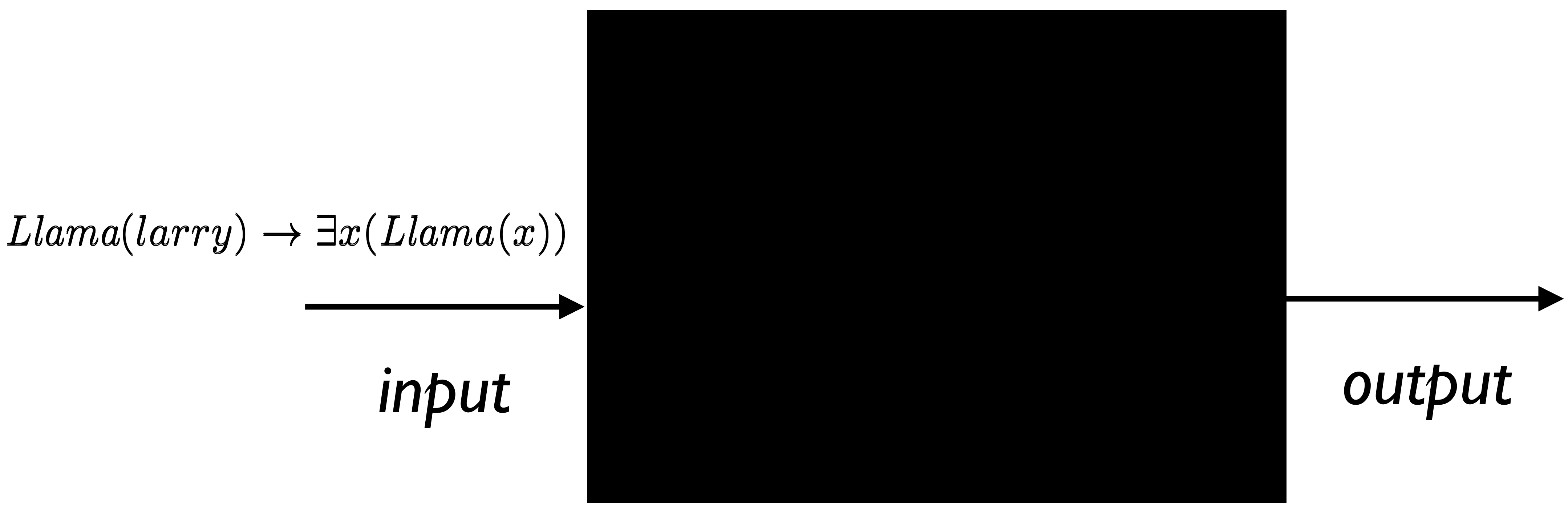
Hard!! — for apparently no polynomial-time algorithm for this!

And now, the Theoremhood Decision Problem, i.e., the *Entscheidungsproblem*, ($\text{THEOREM}_{\text{FOL}}$) for First-Order Logic (FOL)

And now, the Theoremhood Decision Problem, i.e., the *Entscheidungsproblem*, (THEOREM$_{FOL}$) for First-Order Logic (FOL)



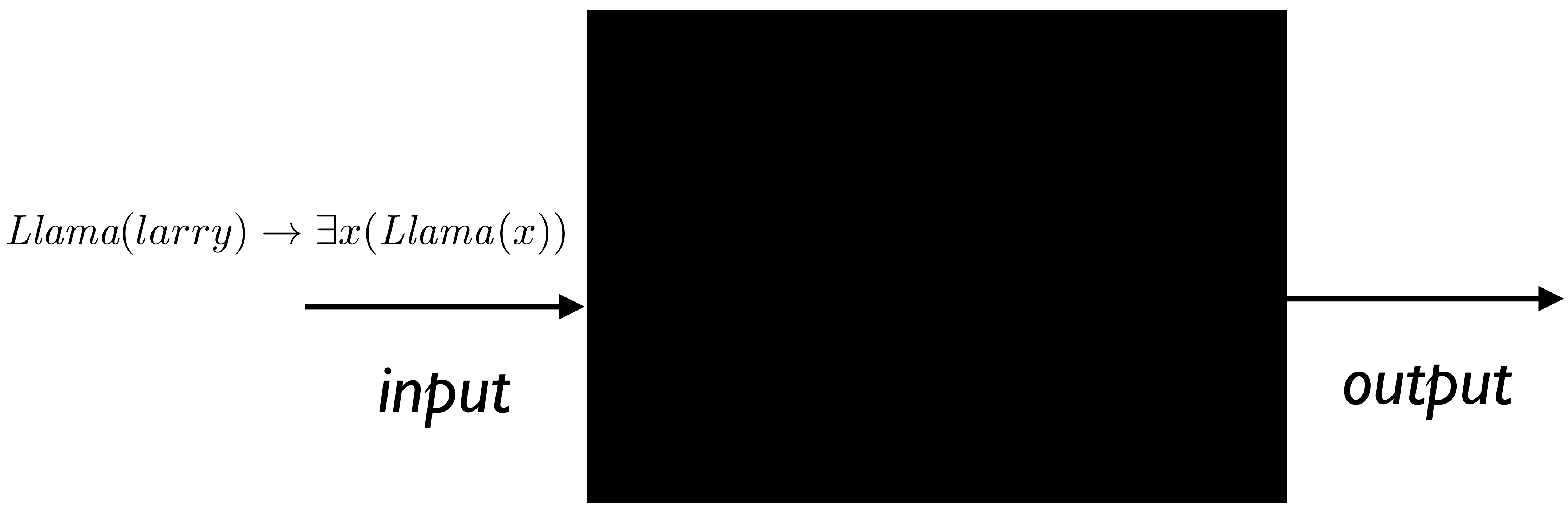
And now, the Theoremhood Decision Problem, i.e., the *Entscheidungsproblem*, ($\text{THEOREM}_{\text{FOL}}$) for First-Order Logic (FOL)



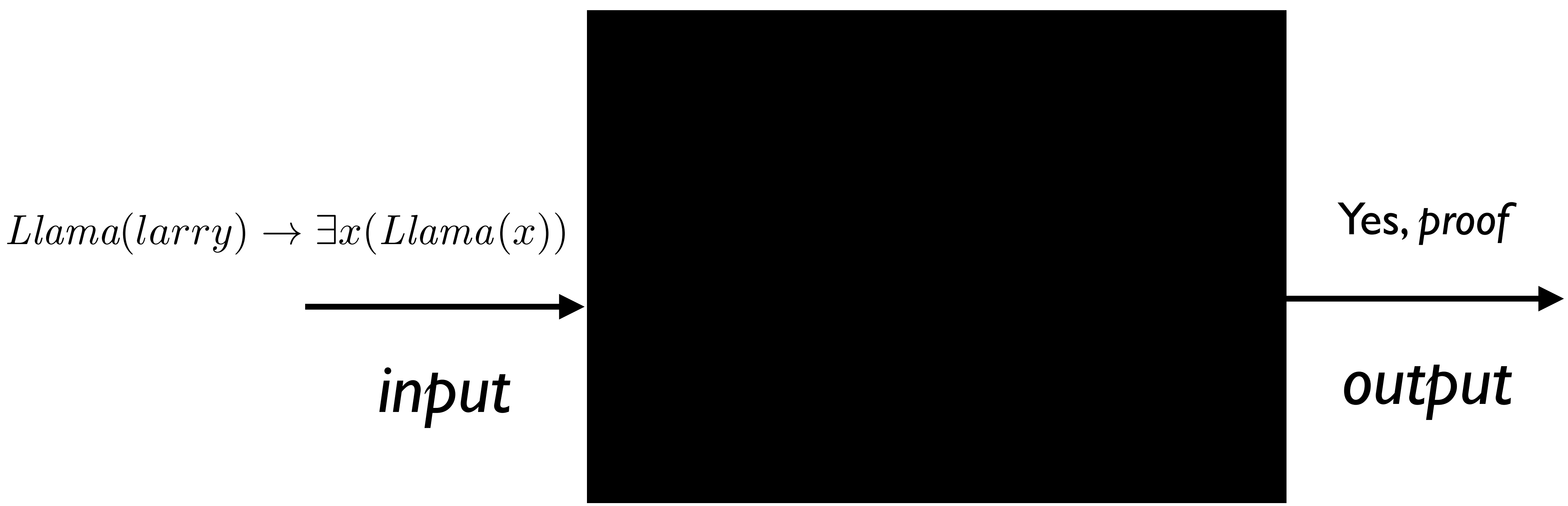
And now, the Theoremhood Decision Problem, i.e., the *Entscheidungsproblem*, ($\text{THEOREM}_{\text{FOL}}$) for First-Order Logic (FOL)

$$Llama(larry) \rightarrow \exists x(Llama(x))$$

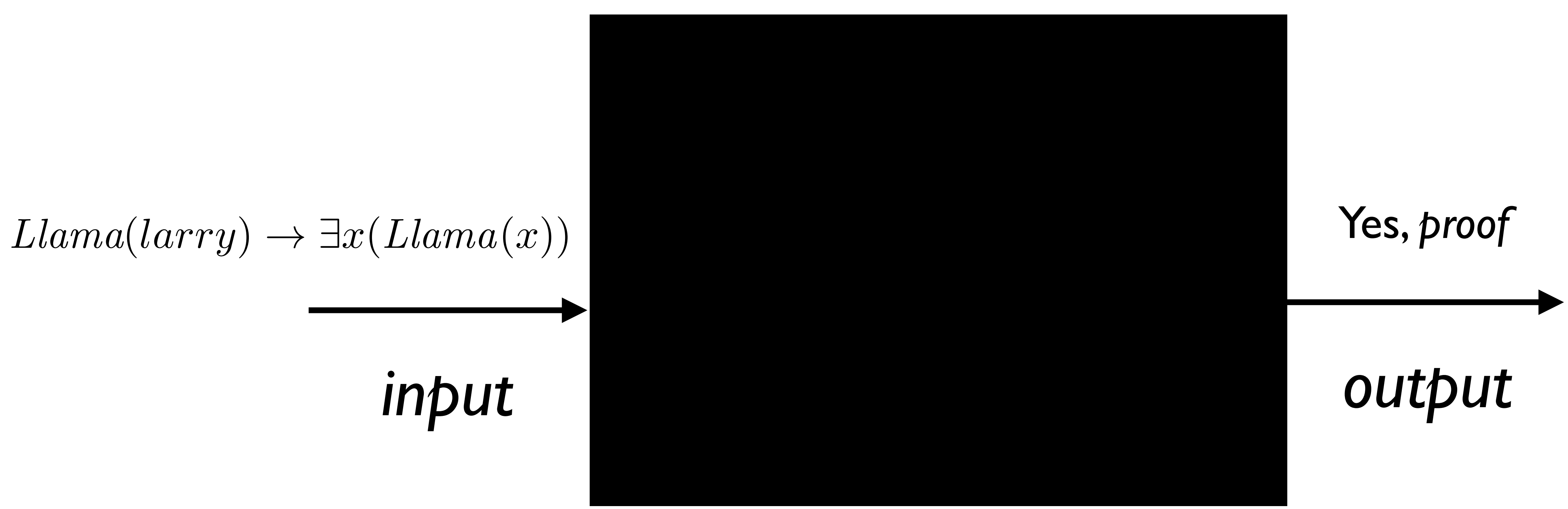
And now, the Theoremhood Decision Problem, i.e., the *Entscheidungsproblem*, ($\text{THEOREM}_{\text{FOL}}$) for First-Order Logic (FOL)



And now, the Theoremhood Decision Problem, i.e., the *Entscheidungsproblem*, (THEOREM$_{FOL}$) for First-Order Logic (FOL)



And now, the Theoremhood Decision Problem, i.e., the *Entscheidungsproblem*, (THEOREM$_{\text{FOL}}$) for First-Order Logic (FOL)



Not just hard: *impossible* for a (and this needed to be *invented* in the course of clarifying and solving the problem) standard computing machine.

Applying this to …
The Singularity Question

Applying this to …
The Singularity Question

$\mathcal{A}$:

	Premise 1	There will be AI (created by HI and such that AI = HI).
	Premise 2	If there is AI, there will be AI^{+} (created by AI).
	Premise 3	If there is AI^{+}, there will be AI^{++} (created by AI^{+}).
$\therefore$	**S**	There will be AI^{++} (= $\mathcal{S}$ will occur).

(Good-Chalmers Argument)

(Kurzweil is an "extrapolationist.")

Applying this to …
The Singularity Question

So, these super-smart machines that will be built by human-level-smart machines, they can't *possibly* be smart enough to solve the *Entscheidungsproblem*. Hence they'll be just (recursively) faster at solving problems we can routinely solve? What's so super-smart about *that*?

LAMA–BDLA

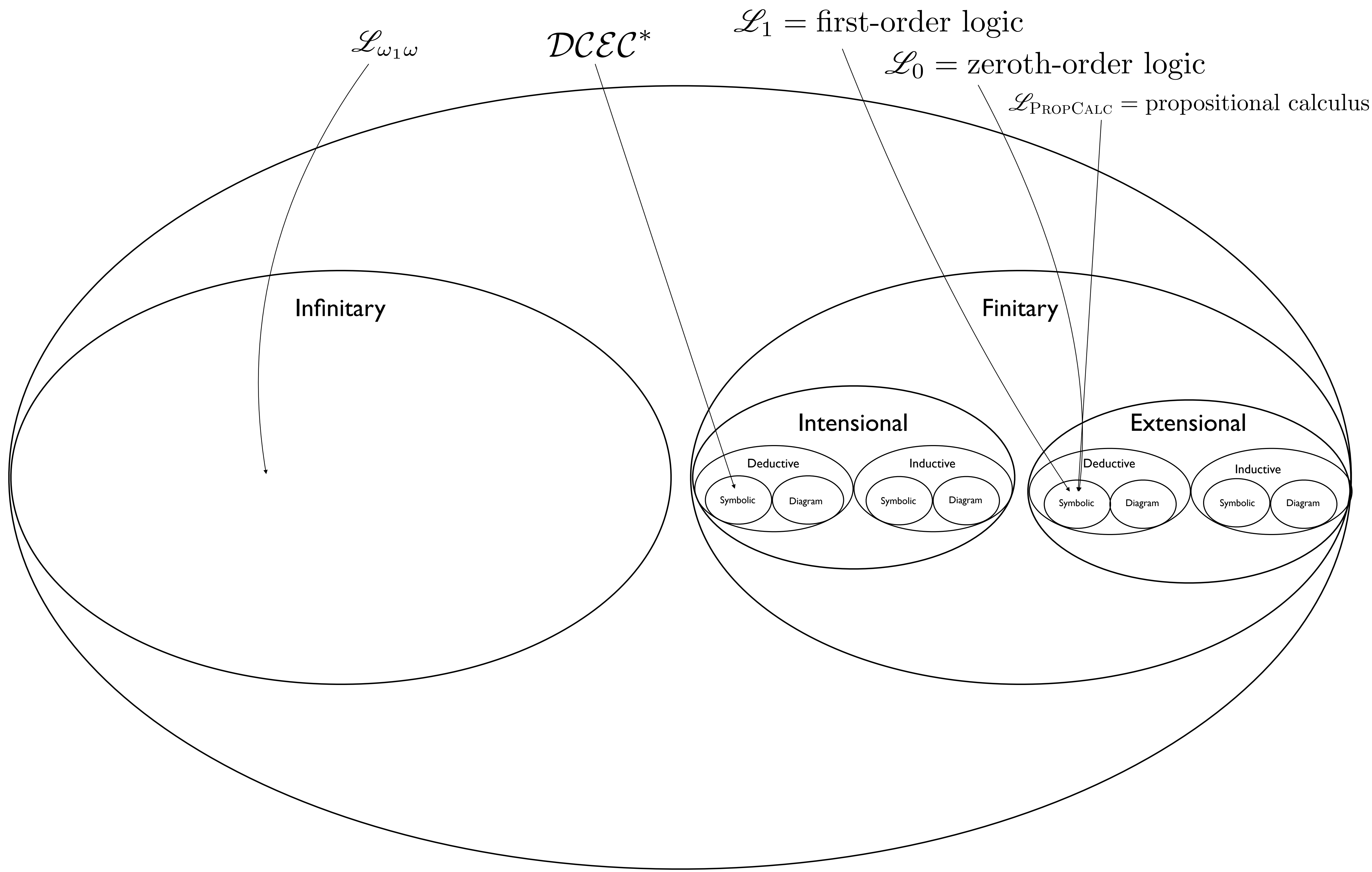
LAMA–BDLA

LAMA–BDLA

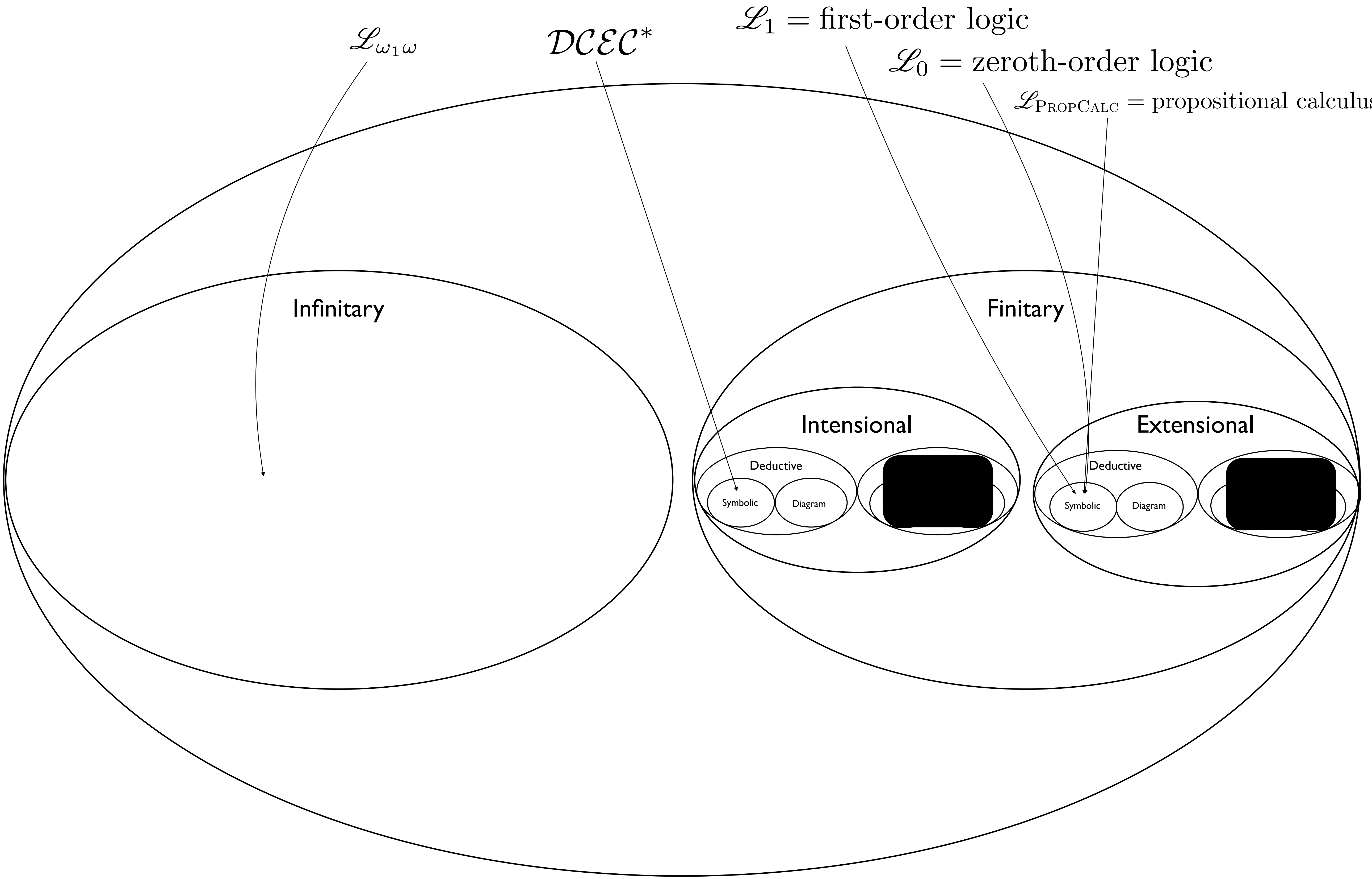
LAMA–BDLA

LAMA–BIL, a bit.

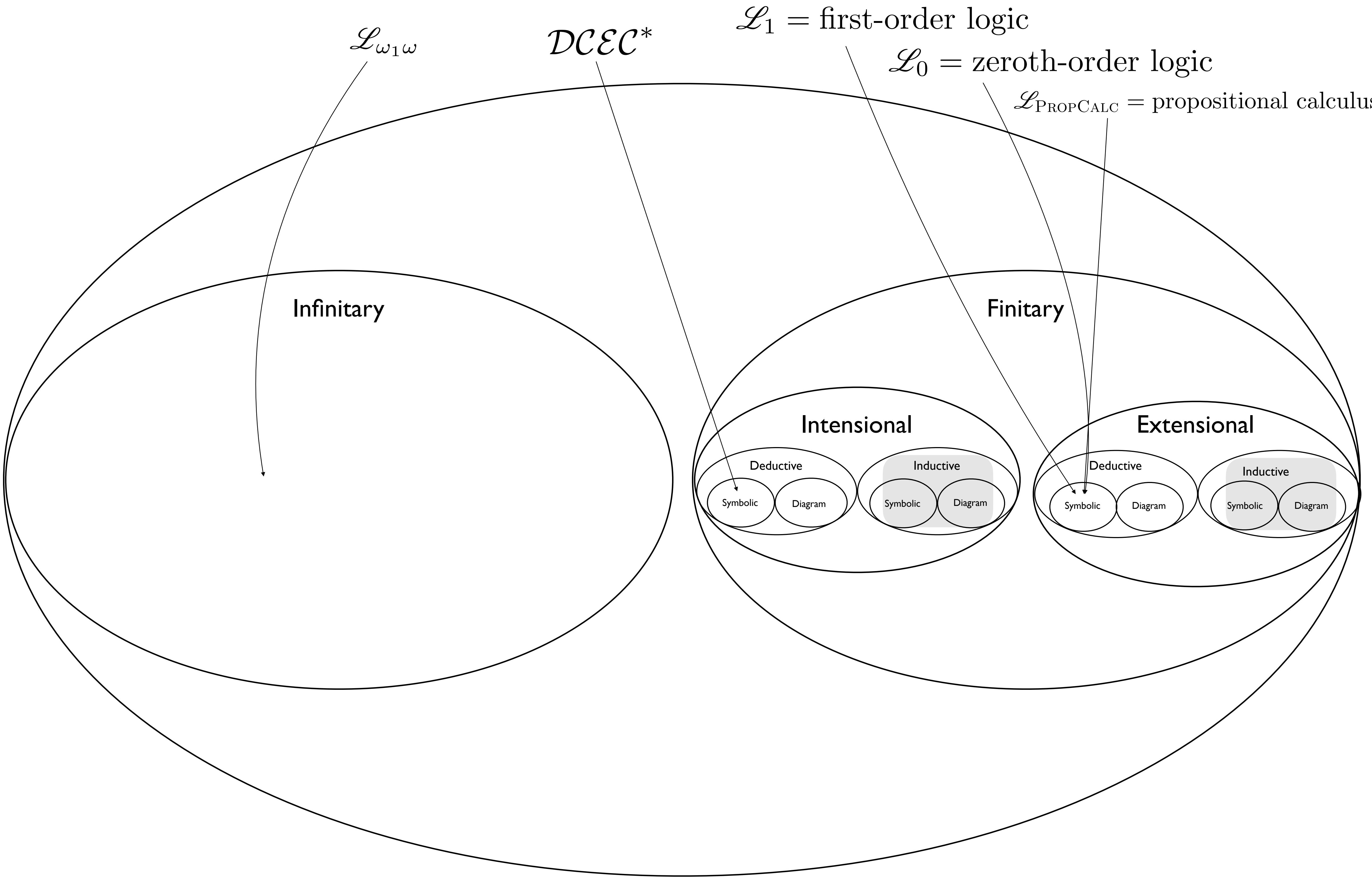
The Universe of Logics



The Universe of Logics



The Universe of Logics

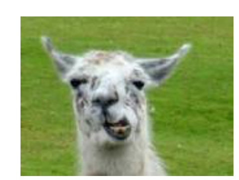


The Monty Hall Problem

$1M

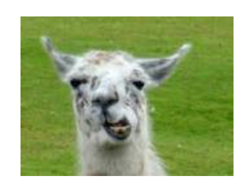
The Monty Hall Problem

$1M



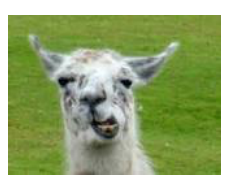
The Monty Hall Problem

$1M

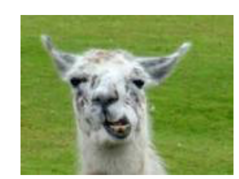


The Monty Hall Problem

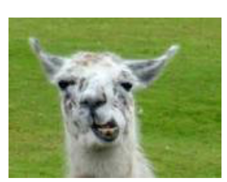
  



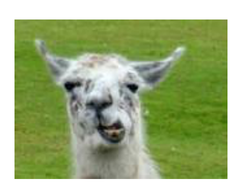
$1M



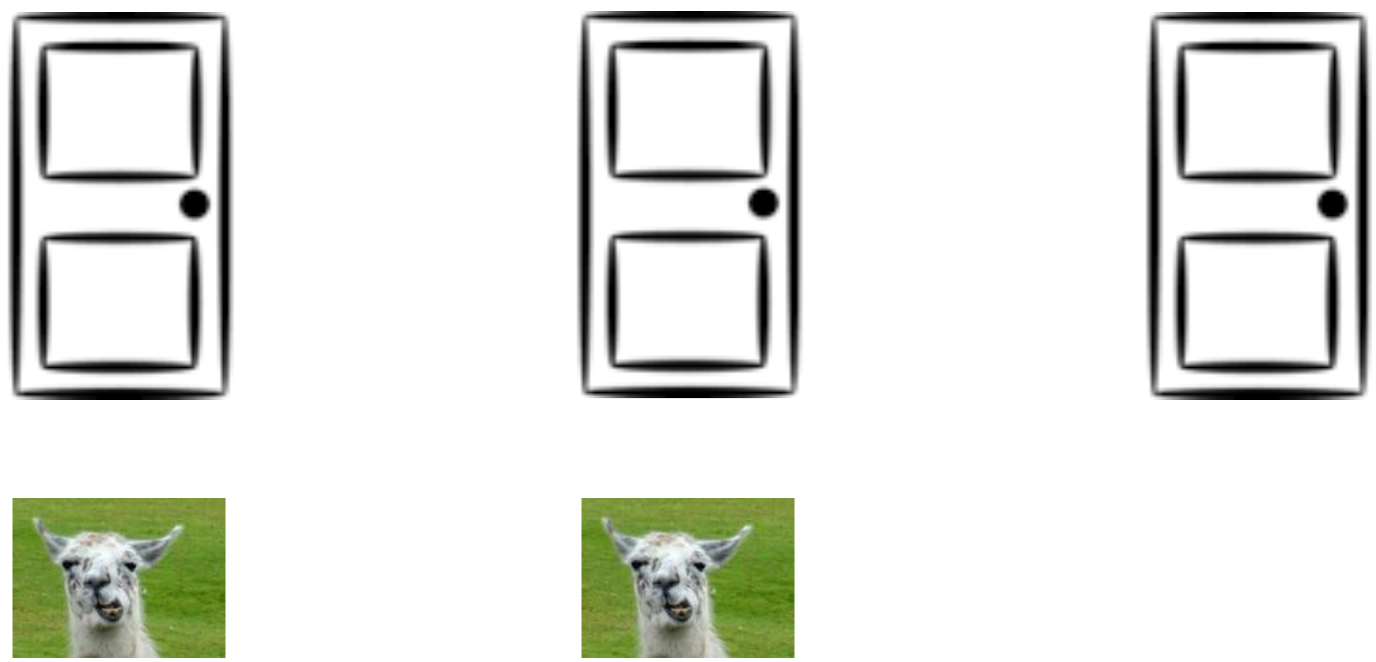
The Monty Hall Problem



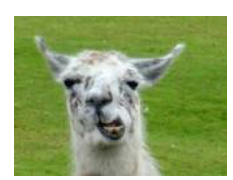
$1M



The Monty Hall Problem

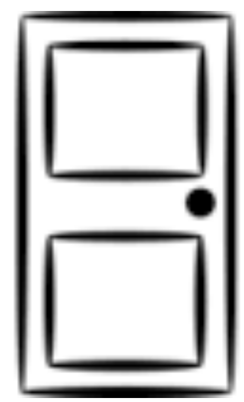


The Monty Hall Problem

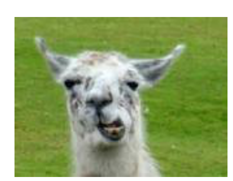


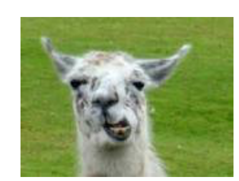
$1M

The Monty Hall Problem



$1M





MHP Defined

Jones has come to a game show, and finds himself thereon selected to play a game on national TV with the show's suave host, Full Monty. Jones is told correctly by Full that hidden behind one of three closed, opaque doors facing the two of them is $1,000,000, while behind each of the other two is a feculent, obstreperous llama whose value on the open market is charitably pegged at $1. Full reminds Jones that this is a game, and a fair one, and that if Jones ends up selecting the door with $1M behind it, all that money will indeed be his. (Jones' net worth has nearly been exhausted by his expenditures in traveling to the show.) Full also reminds Jones that he (= Full) knows what's behind each door, fixed in place until the game ends.

Full asks Jones to select which door he wants the contents of. Jones says, "Door 1." Full then says: "Hm. Okay. Part of this game is my revealing at this point what's behind one of the doors you didn't choose. So ... let me show you what's behind Door 3." Door 3 opens to reveal a very unsavory llama. Full now to Jones: "Do you want to switch to Door 2, or stay with Door 1? You'll get what's behind the door of your choice, and our game will end." Full looks briefly into the camera, directly.

(P1.1) What should Jones do if he's rational?

(P1.2) Prove that your answer is correct. (Diagrammatic proofs are allowed.)

(P1.3) A quantitative hedge fund manager with a PhD in finance from Harvard zipped this email off to Full before Jones made his decision re. switching or not: "Switching would be a royal waste of time (and time is money!). Jones hasn't a doggone clue what's behind Door 1 or Door 2, and it's obviously a 50/50 chance to win whether he stands firm or switches. So the chap shouldn't switch!" Is the fund manager right? Prove that your diagnosis is correct.

(P1.4) Can these answers and proofs be exclusively Bayesian in nature?

The Switching Policy Rational!

Proof: Our overarching technique will be proof by cases.

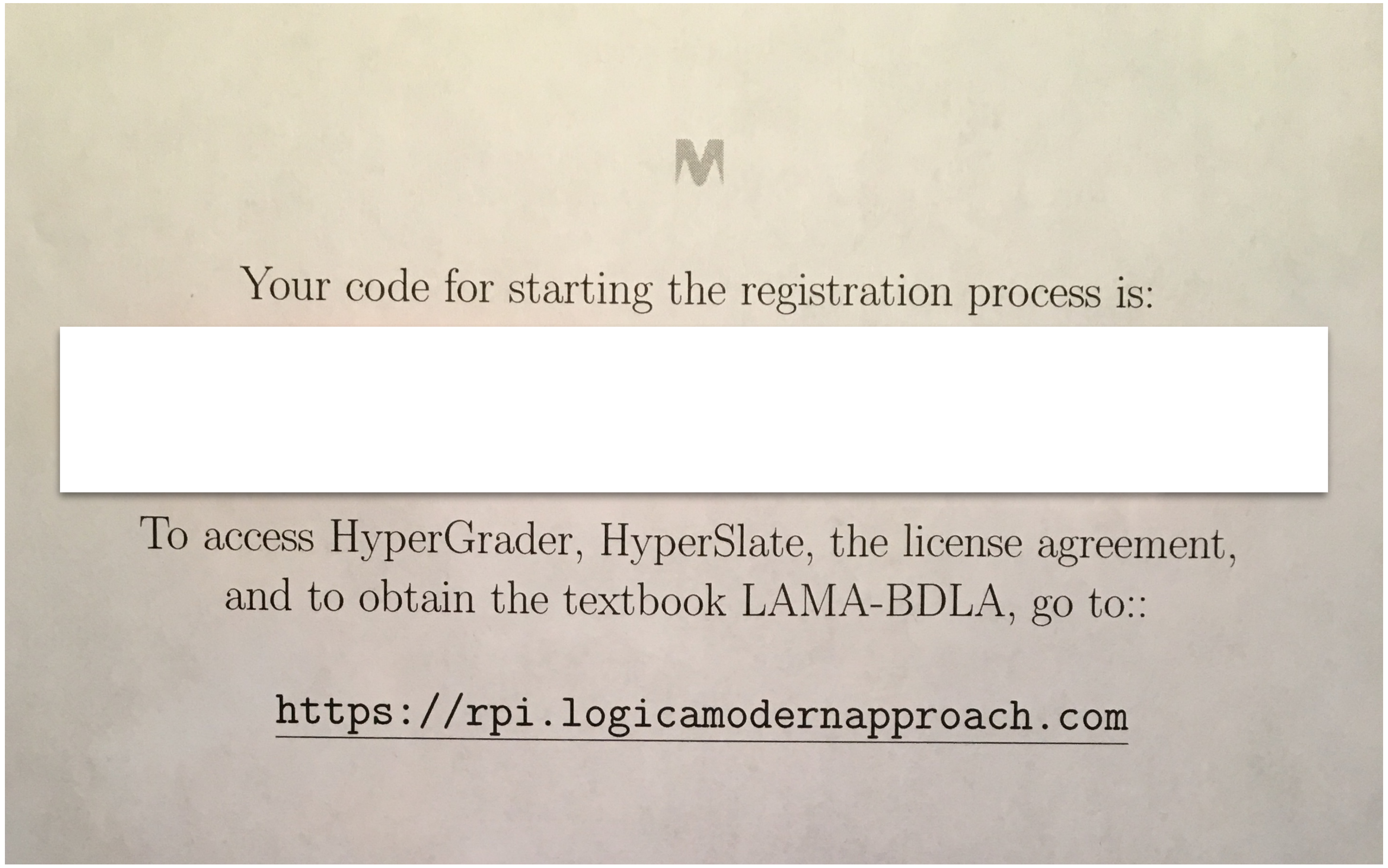
We denote the possible cases for initial distribution using a simple notation, according to which for example 'LLM' means that, there is a lama behind Door 1, a llama behind Door 2, and the million dollars behind Door 3. With this notation in hand, our three starting cases are: Case 1: MLL; Case 2: LML; Case 3: LLM. There are only three top-level cases for distribution. The odds of picking at the start the million-dollar door is 1/3, obviously — for each case. Hence we know that the odds of a HOLD policy winning is 1/3.

Now we proceed in a proof by sub-cases under the three cases above, to show that the overall odds of a SWITCH policy is greater than 1/3. Each sub-case is simply based on what the initial choice by Jones is, under one of the three main cases. Here we go:

Suppose Case 3, LLM, holds, and that [this (Case 3.1) is the first of three sub-cases under Case 3] Jones picks Door 1. Then FM must reveal Door 2 to reveal a llama. Switching to Door 3 wins, guaranteed. In sub-case 3.2 suppose that J's choice Door 2. Then FM will reveal Door 1. Again, switching to Door 3 wins, guaranteed. In the final sub-case, J initially selects Door 3 under Case 3; this is sub-case 3.3. Here, FM shows either Door 1 or Door 2 (as itself a random choice). This time switching loses, guaranteed. Hence, in two of the sub-cases out of three (2/3), winning is guaranteed (*prob* of 1). An exactly parallel result can be deduced for Case 2 and Case 1; i.e., in each of these two, in two of the three (2/3) sub-cases winning is 1. Hence the odds of winning by following the switching policy is 2/3, which is greater than 1/3. Hence it's rational to be a switcher. **QED**

Logistics …

The Starting Code to Purchase in Bookstore

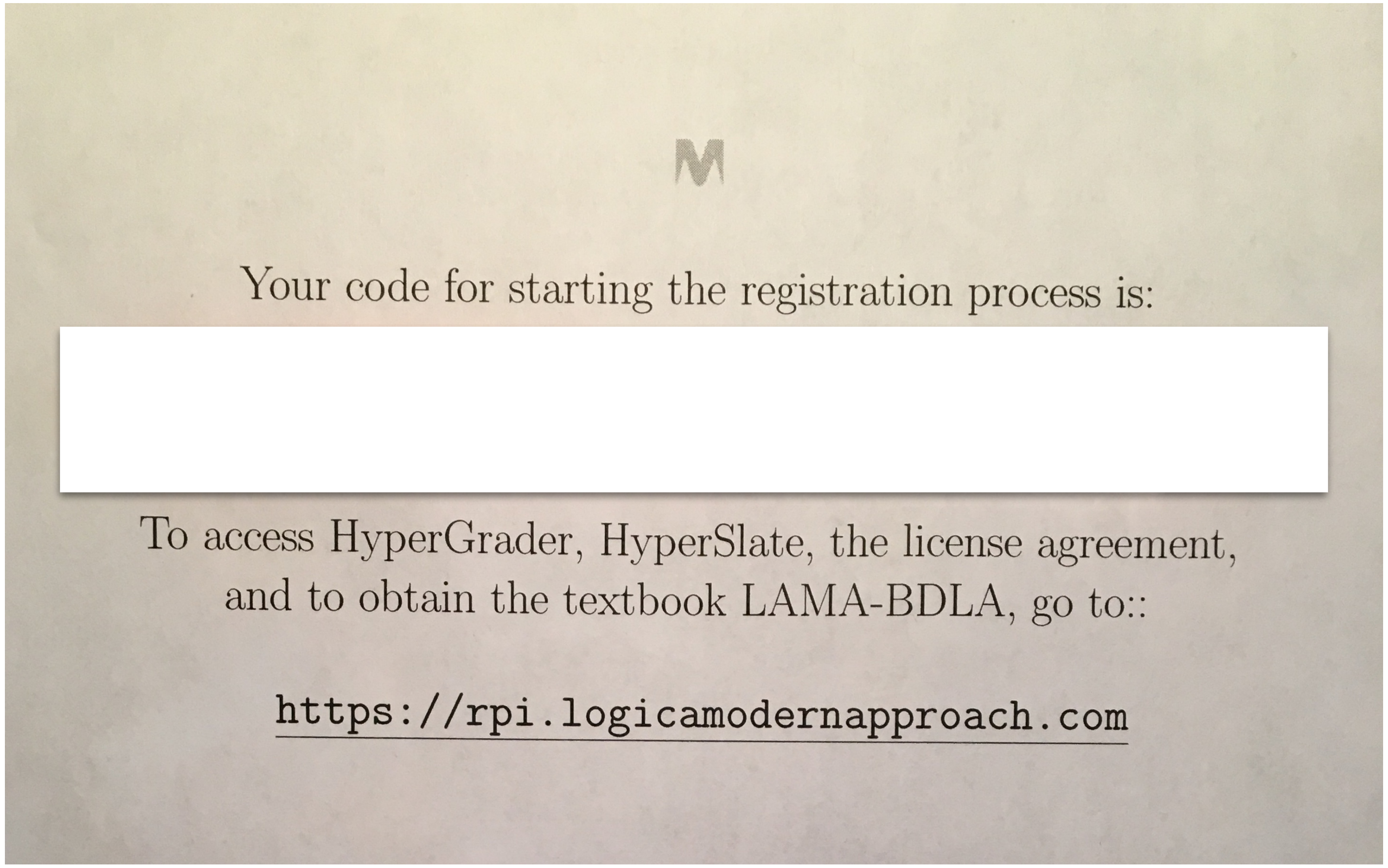


Your code for starting the registration process is:

To access HyperGrader, HyperSlate, the license agreement, and to obtain the textbook LAMA-BDLA, go to::

https://rpi.logicamodernapproach.com

The Starting Code to Purchase in Bookstore



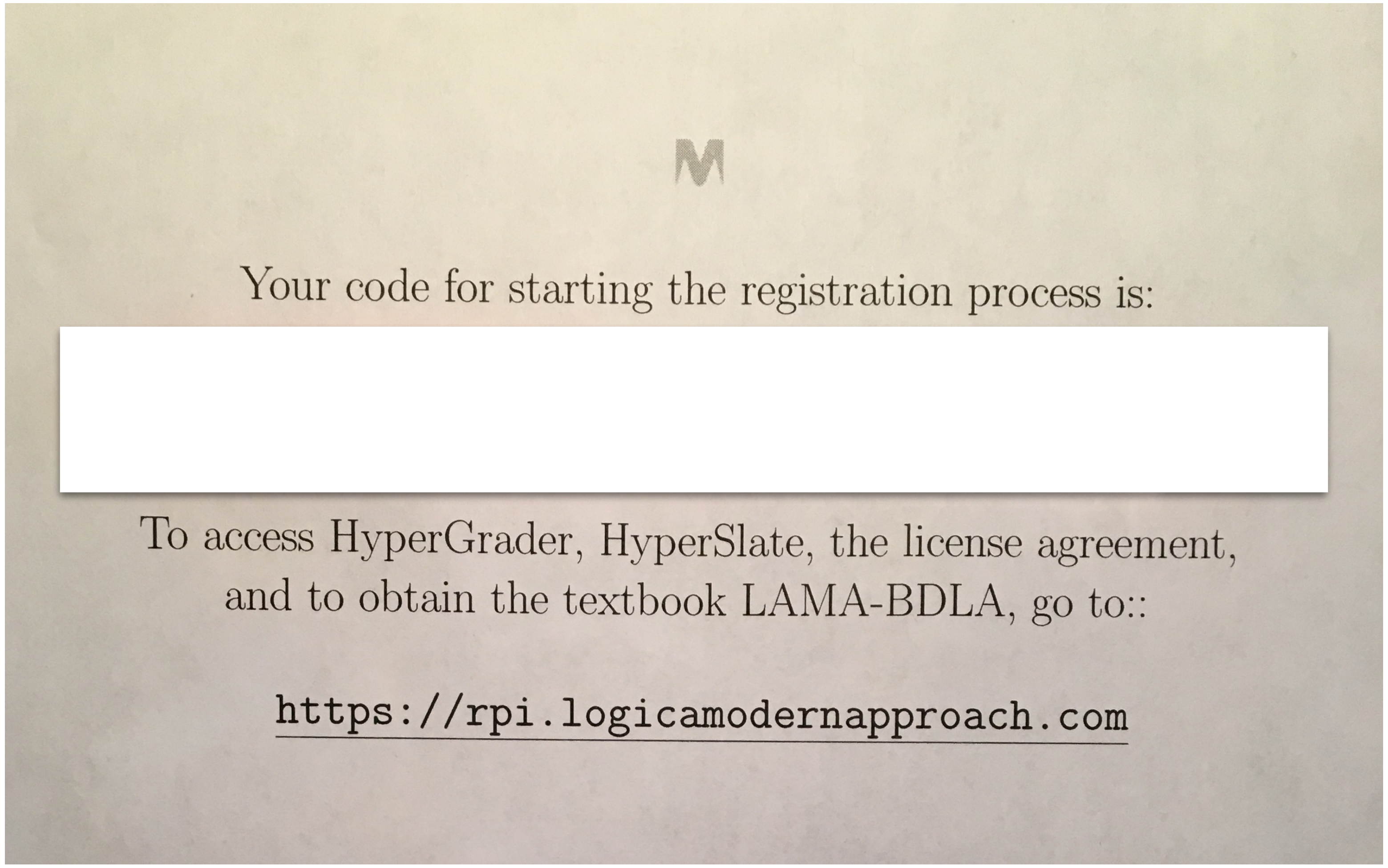
Your code for starting the registration process is:

To access HyperGrader, HyperSlate, the license agreement, and to obtain the textbook LAMA-BDLA, go to::

https://rpi.logicamodernapproach.com

Once seal broken on envelope, no return. Remember from first class, any reservations, opt for "Stanford" paradigm, with its software instead of LAMA® paradigm!

The Starting Code to Purchase in Bookstore


Your code for starting the registration process is:

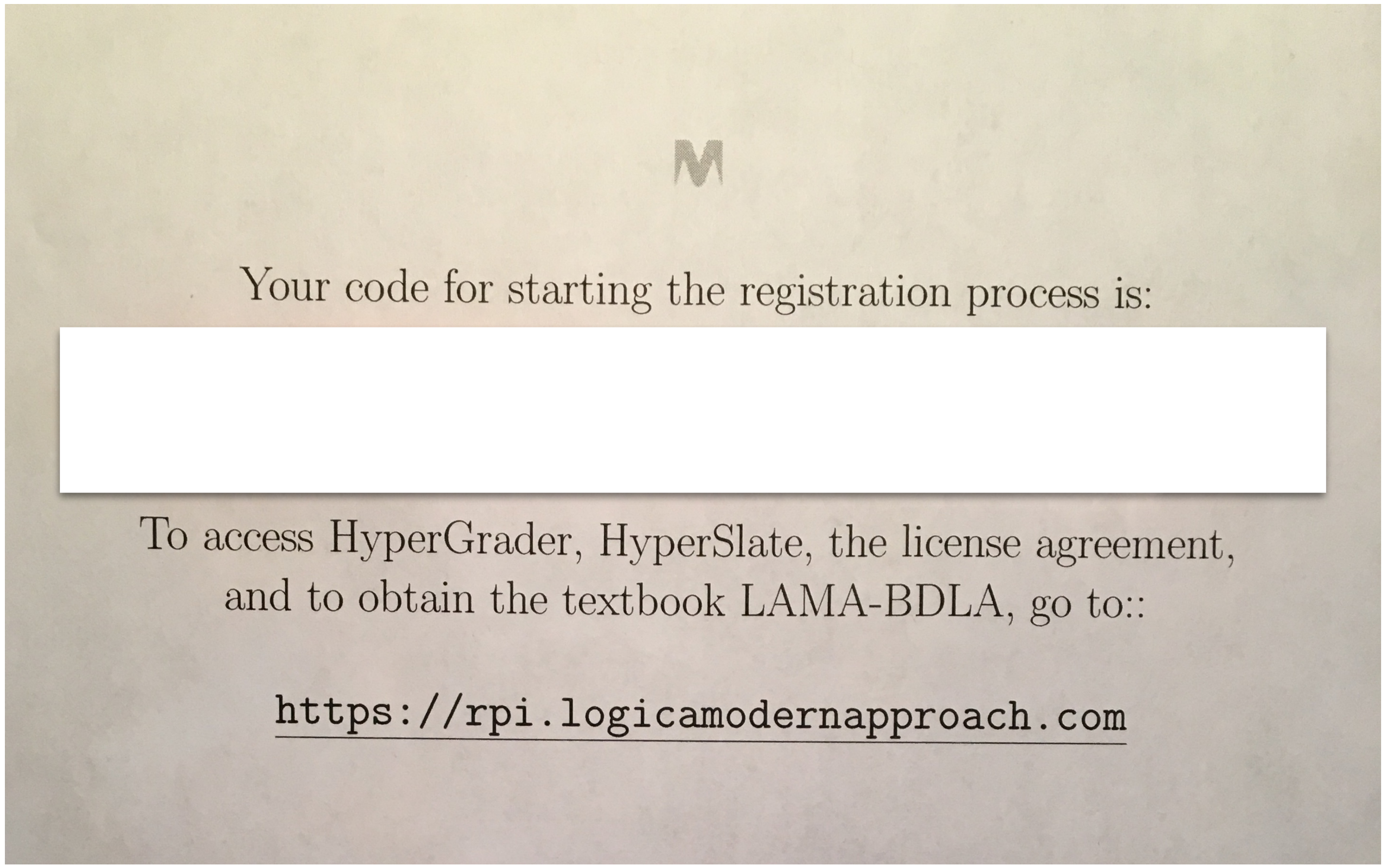
To access HyperGrader, HyperSlate, the license agreement, and to obtain the textbook LAMA-BDLA, go to::

https://rpi.logicamodernapproach.com

Once seal broken on envelope, no return. Remember from first class, any reservations, opt for "Stanford" paradigm, with its software instead of LAMA® paradigm!

The email address you enter is case-sensitive!

The Starting Code to Purchase in Bookstore


Your code for starting the registration process is:

To access HyperGrader, HyperSlate, the license agreement, and to obtain the textbook LAMA-BDLA, go to::

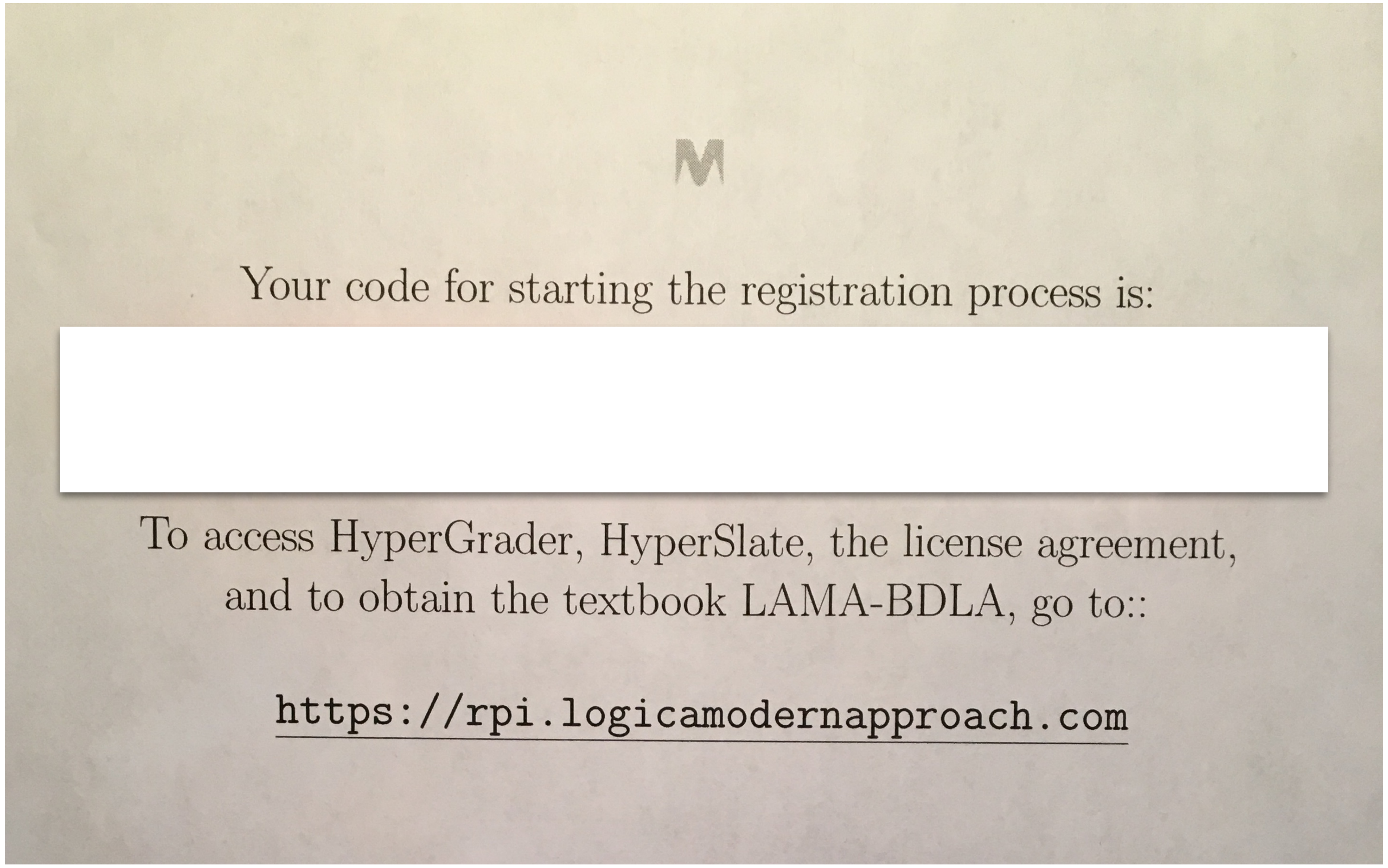
https://rpi.logicamodernapproach.com

Once seal broken on envelope, no return. Remember from first class, any reservations, opt for "Stanford" paradigm, with its software instead of LAMA® paradigm!

The email address you enter is case-sensitive!

Your OS and browser must be fully up-to-date; Chrome is the best choice, browser-wise (though I use Safari).

The Starting Code to Purchase in Bookstore



Your code for starting the registration process is:

To access HyperGrader, HyperSlate, the license agreement, and to obtain the textbook LAMA-BDLA, go to::

https://rpi.logicamodernapproach.com

Once seal broken on envelope, no return. Remember from first class, any reservations, opt for "Stanford" paradigm, with its software instead of LAMA® paradigm!

The email address you enter is case-sensitive!

Your OS and browser must be fully up-to-date; Chrome is the best choice, browser-wise (though I use Safari).

Watch that the link emailed to you doesn't end up being classified as spam.

Introduction to (Formal) Logic (and AI)

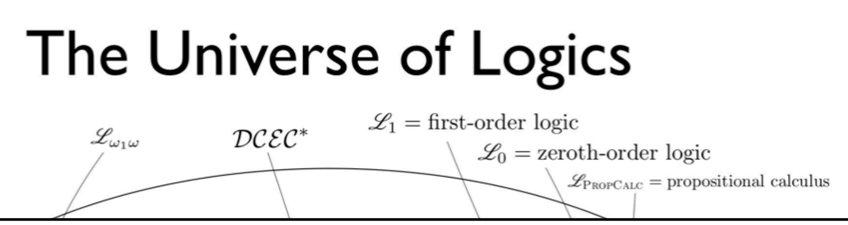
Spring 2023 Edition of IFLAI1 (“eye” • “fly” • “one”)

Selmer Bringsjord

Table of Contents

Learning logic with patented AI technology.

with Naveen Sundar G.
∧ KB Foushée ∧ \(\ldots\)



Logikk kan gi dyp glede!