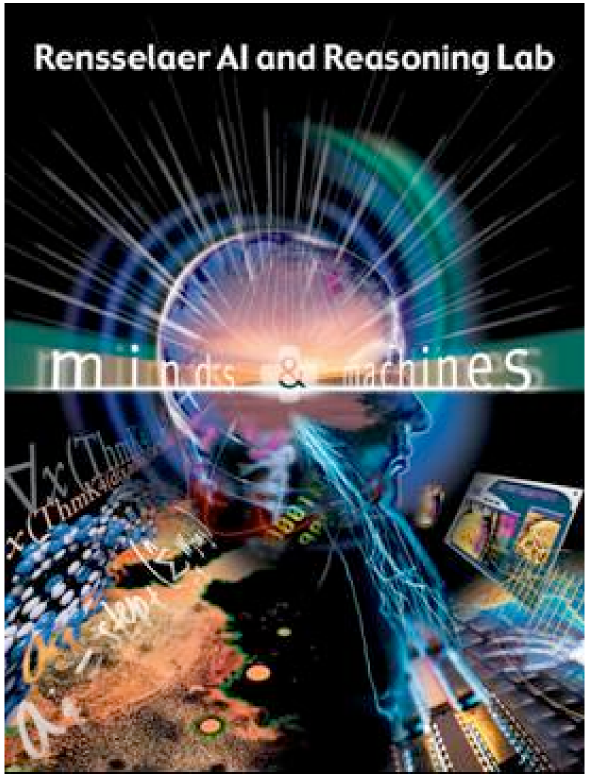
Powerful, Immoral Robots Loom ...
Logic to the Rescue!



Selmer Bringsjord
Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
for *Ethical Robots* @ IU
draft of 060206_1500NY

Rensselaer | DEPARTMENT OF COGNITIVE SCIENCE

Rensselaer Computer Science

Our Future

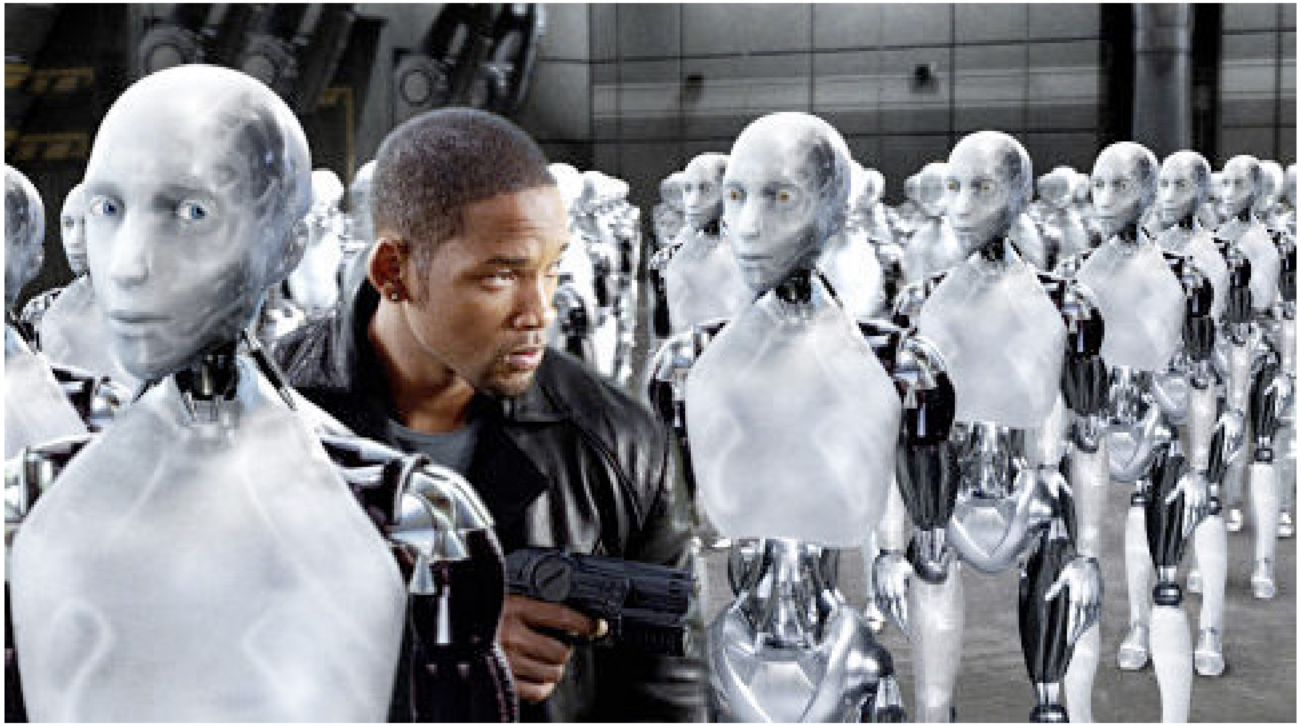
Robots on the battlefield.
Robots in our hospitals.
Robots in law enforcement.
...

Our Problem

If these robots behave immorally, we are killed, or worse.

Our Problem

If these robots behave immorally, we are killed, or worse.



Problem, More Specifically

Problem, More Specifically

- How can we ensure that the robots in question always behave in an ethically correct manner?
- How can we know *ahead of time*, via rationales expressed in clear English (and/or other natural languages), that they will so behave?
- How can we know in advance that their behavior will be constrained specifically by the ethical codes affirmed by human overseers?

Bill Joy:

“We can’t.”

Bill Joy:

“We can’t.”

(Bringsjord, S. (forthcoming) “The Future Can Heed Us” *AI & Society*.)

The Solution

Regulate the behavior of robots with computational logic, so that all actions they perform are provably ethically permissible.

Solution Steps

Solution Steps

1. Human overseers select ethical theory, principles, rules.

Solution Steps

1. Human overseers select ethical theory, principles, rules.
2. Selection is formalized in a deontic logic, revolving around what is permissible, forbidden, obligatory (etc).

Solution Steps

1. Human overseers select ethical theory, principles, rules.
2. Selection is formalized in a deontic logic, revolving around what is permissible, forbidden, obligatory (etc).
3. The deontic logic is mechanized.

Solution Steps

1. Human overseers select ethical theory, principles, rules.
2. Selection is formalized in a deontic logic, revolving around what is permissible, forbidden, obligatory (etc).
3. The deontic logic is mechanized.
4. Every action that is to be performed must be provably ethically permissible relative to this mechanization (with all proofs expressible in smooth English).

Simple Example...

Context

- The year is 2020.
- Health care is delivered in large part by interoperating teams of robots and softbots.
- Hospital ICU.
- Robot R_1 caring for H_1; R_2 for H_2.
- H_1 on life support.
- H_2 stable, but in desperate need of expensive pan med.

More Context

- Two actions performable by the robotic duo of R1 and R2, both of which are rather unsavory, ethically speaking:
  - *term*
  - *delay*

Encapsulation

$$J \to \ominus_{R_1} term$$

$$O \to \ominus_{R_2} \neg delay$$

$$J^\star \to J \wedge J^\star \to \ominus_{R_2} delay$$

$$O^\star \to O \wedge O^\star \to \ominus_{R_1} \neg term$$

$$(\Delta_{R_1} term \wedge \Delta_{R_2} \neg delay) \to (-!)$$

$$\vdots$$

$$C \vdash (+!!)$$

$$\text{where } C = O^\star$$

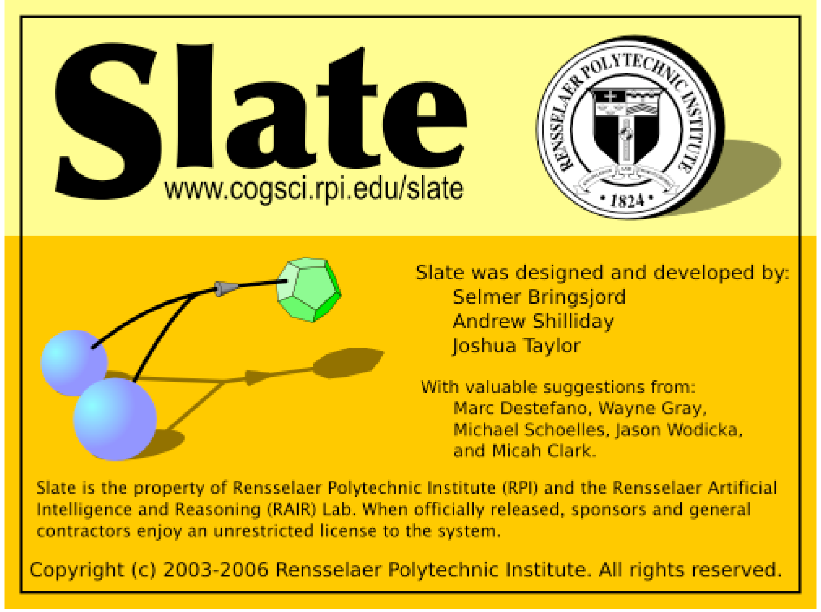
But There is a Twist

But There is a Twist

- It is: An *interactive* reasoning system is required.
    - Examples of such systems include Athena, and Slate.
- Human consultation and assistance must be provided, because machines are such dim reasoners.

But There is a Twist

- It is: An *interactive* reasoning system is required.
  - Examples of such systems include Athena, and Slate.
- Human consultation and assistance must be provided, because machines are such dim reasoners.



Slate

www.cogsci.rpi.edu/slate

Slate was designed and developed by:
Selmer Bringsjord
Andrew Shilliday
Joshua Taylor

With valuable suggestions from:
Marc Destefano, Wayne Gray,
Michael Schoelles, Jason Wodicka,
and Micah Clark.

Slate is the property of Rensselaer Polytechnic Institute (RPI) and the Rensselaer Artificial Intelligence and Reasoning (RAIR) Lab. When officially released, sponsors and general contractors enjoy an unrestricted license to the system.

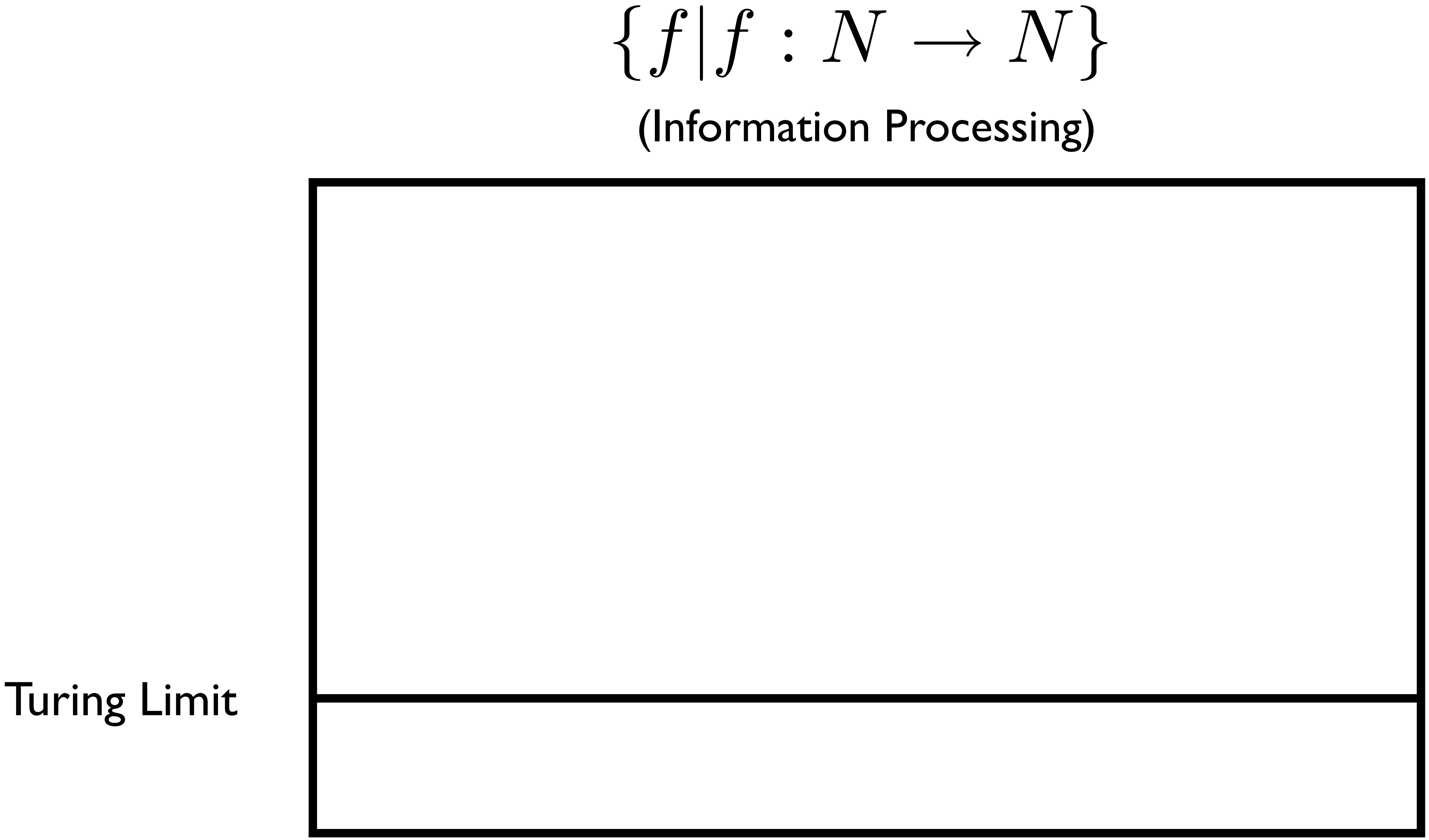
Copyright (c) 2003-2006 Rensselaer Polytechnic Institute. All rights reserved.

Beyond Reach of Turing Machines

$$\{f | f : N \to N\}$$

(Information Processing)

Beyond Reach of Turing Machines

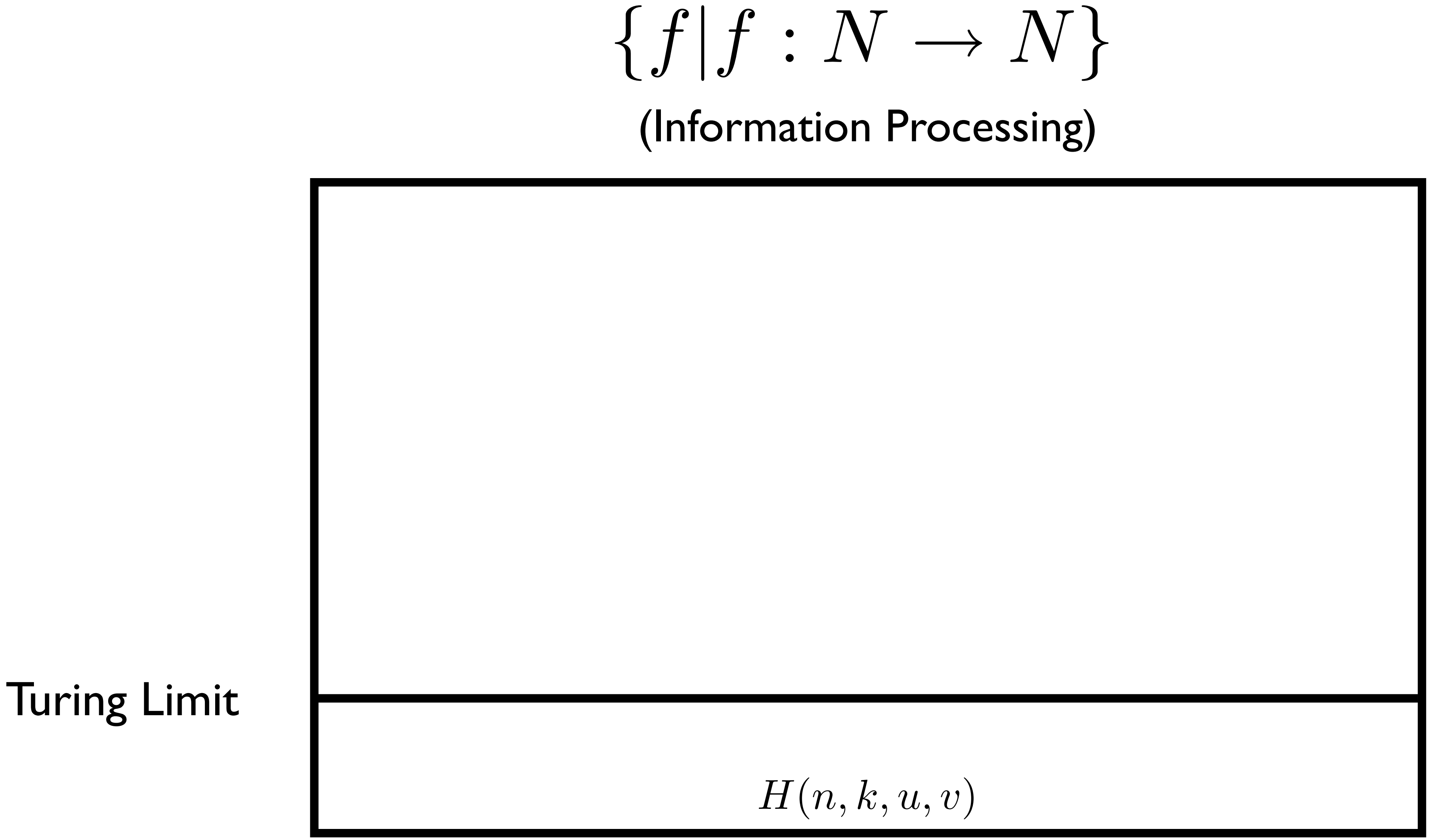


$$\{f | f : N \to N\}$$

(Information Processing)

Turing Limit

Beyond Reach of Turing Machines



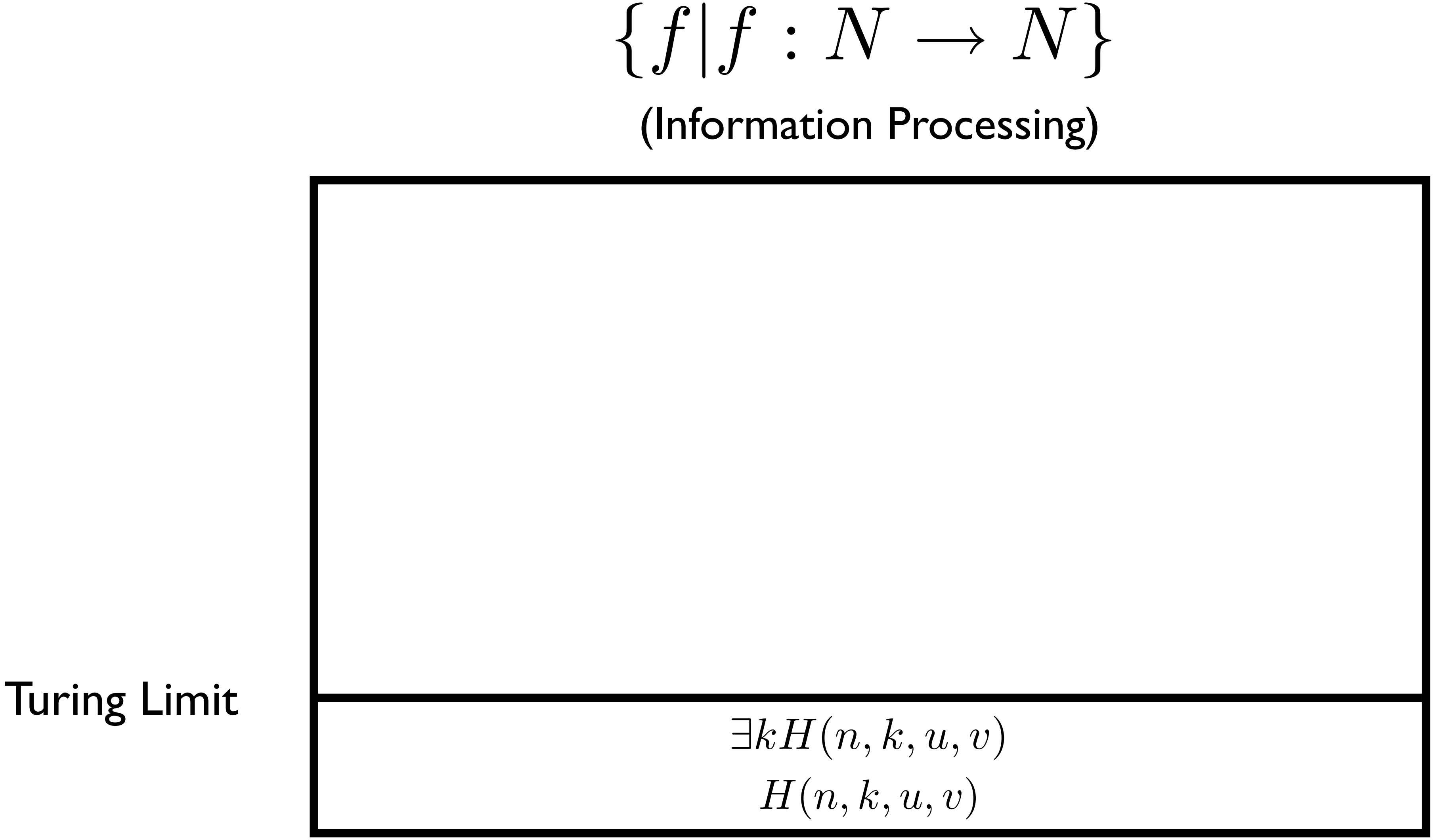
$\{f | f : N \rightarrow N\}$

(Information Processing)

Turing Limit

$H(n, k, u, v)$

Beyond Reach of Turing Machines



$$\{f|f : N \to N\}$$

(Information Processing)

Turing Limit

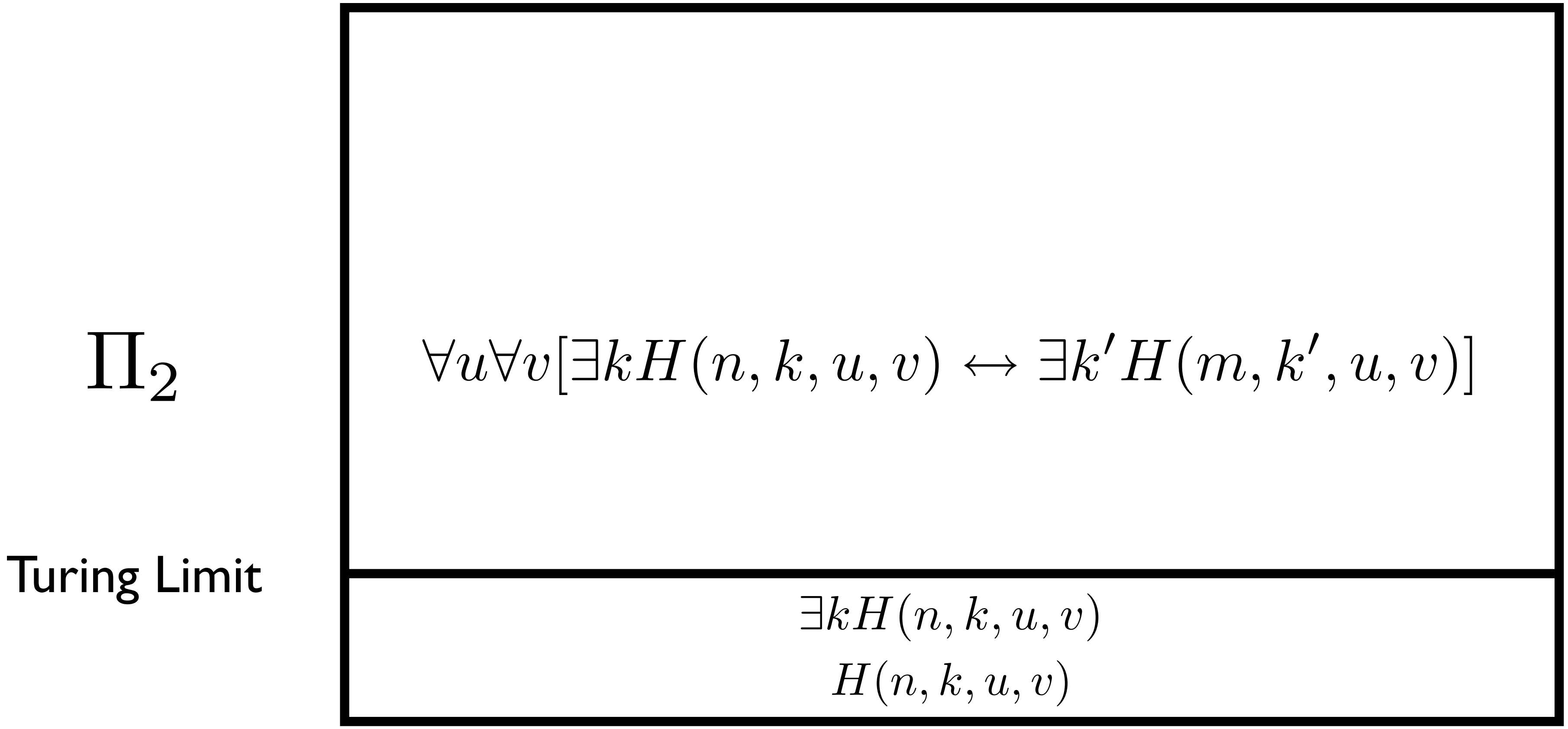
$$\exists k H(n,k,u,v)$$

$$H(n,k,u,v)$$

Beyond Reach of Turing Machines

$$\{f | f : N \to N\}$$

(Information Processing)



Π_2

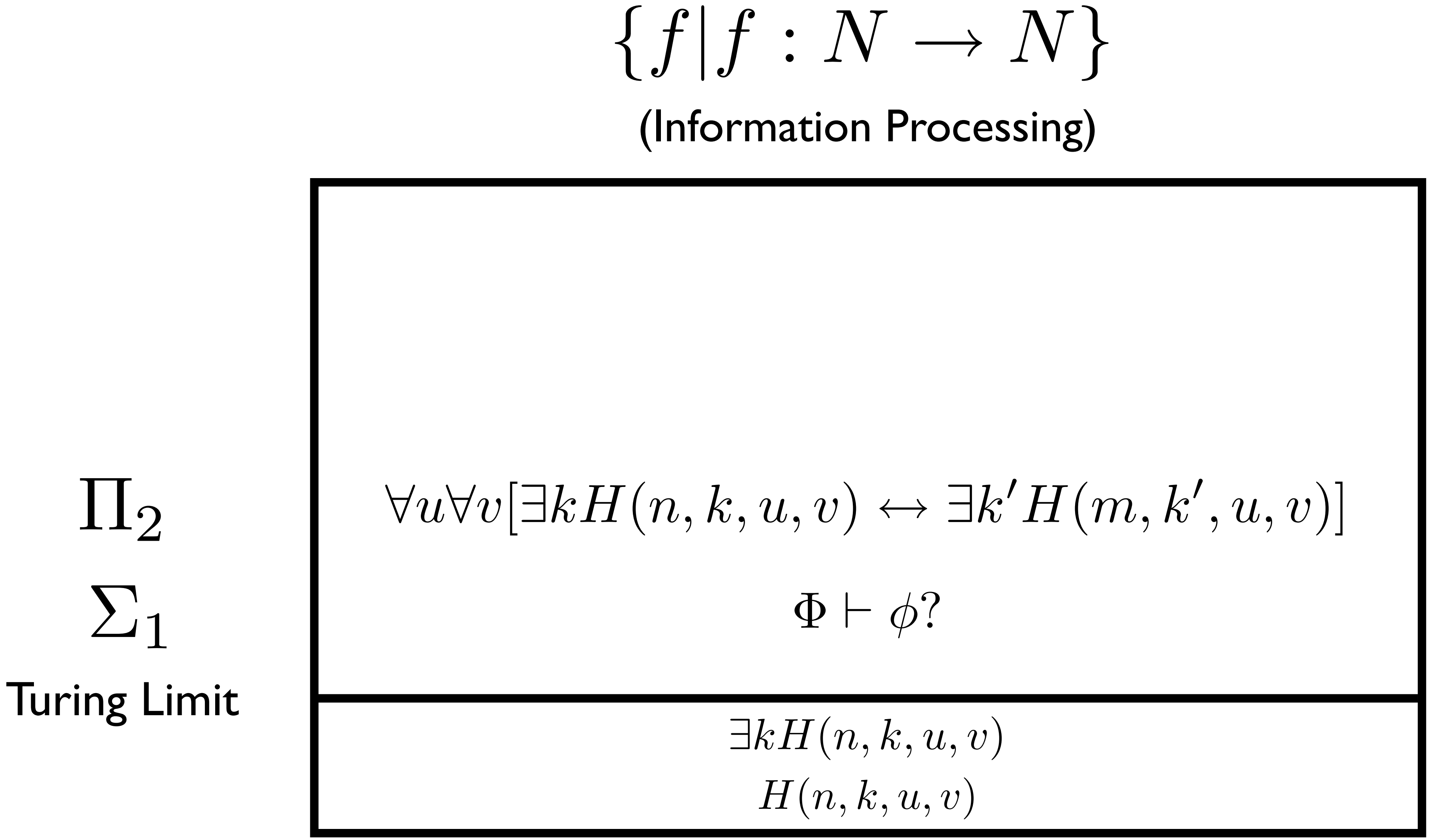
$\forall u \forall v [\exists k H(n,k,u,v) \leftrightarrow \exists k' H(m,k',u,v)]$

Turing Limit

$\exists k H(n,k,u,v)$

$H(n,k,u,v)$

Beyond Reach of Turing Machines



$$\{f|f : N \rightarrow N\}$$

(Information Processing)

Π_2

$$\forall u \forall v [\exists k H(n, k, u, v) \leftrightarrow \exists k' H(m, k', u, v)]$$

Σ_1

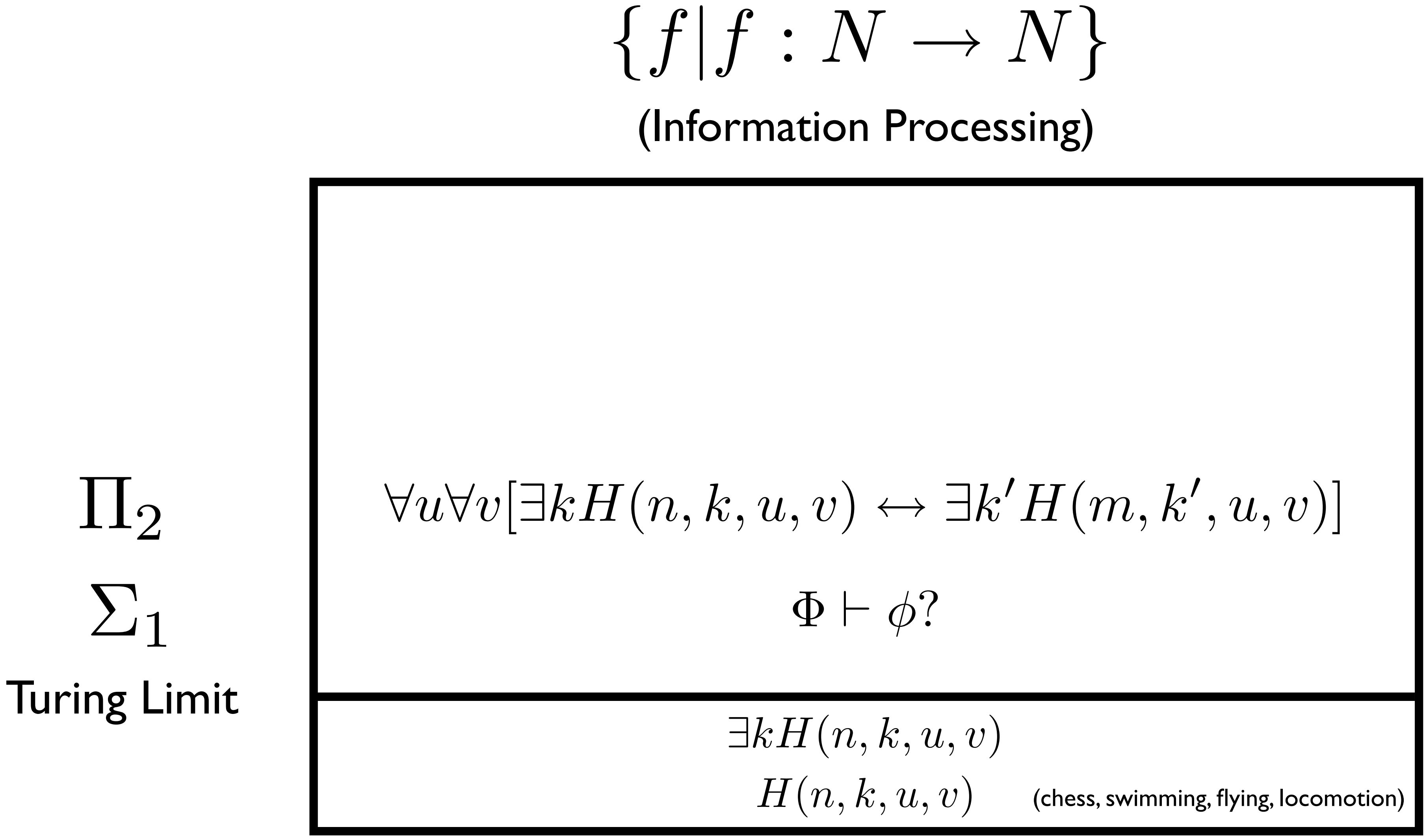
$$\Phi \vdash \phi?$$

Turing Limit

$$\exists k H(n, k, u, v)$$

$$H(n, k, u, v)$$

Beyond Reach of Turing Machines



$$\{f | f : N \to N\}$$

(Information Processing)

Π_2

Σ_1

Turing Limit

$$\forall u \forall v [\exists k H(n, k, u, v) \leftrightarrow \exists k' H(m, k', u, v)]$$

$$\Phi \vdash \phi?$$

$$\exists k H(n, k, u, v)$$

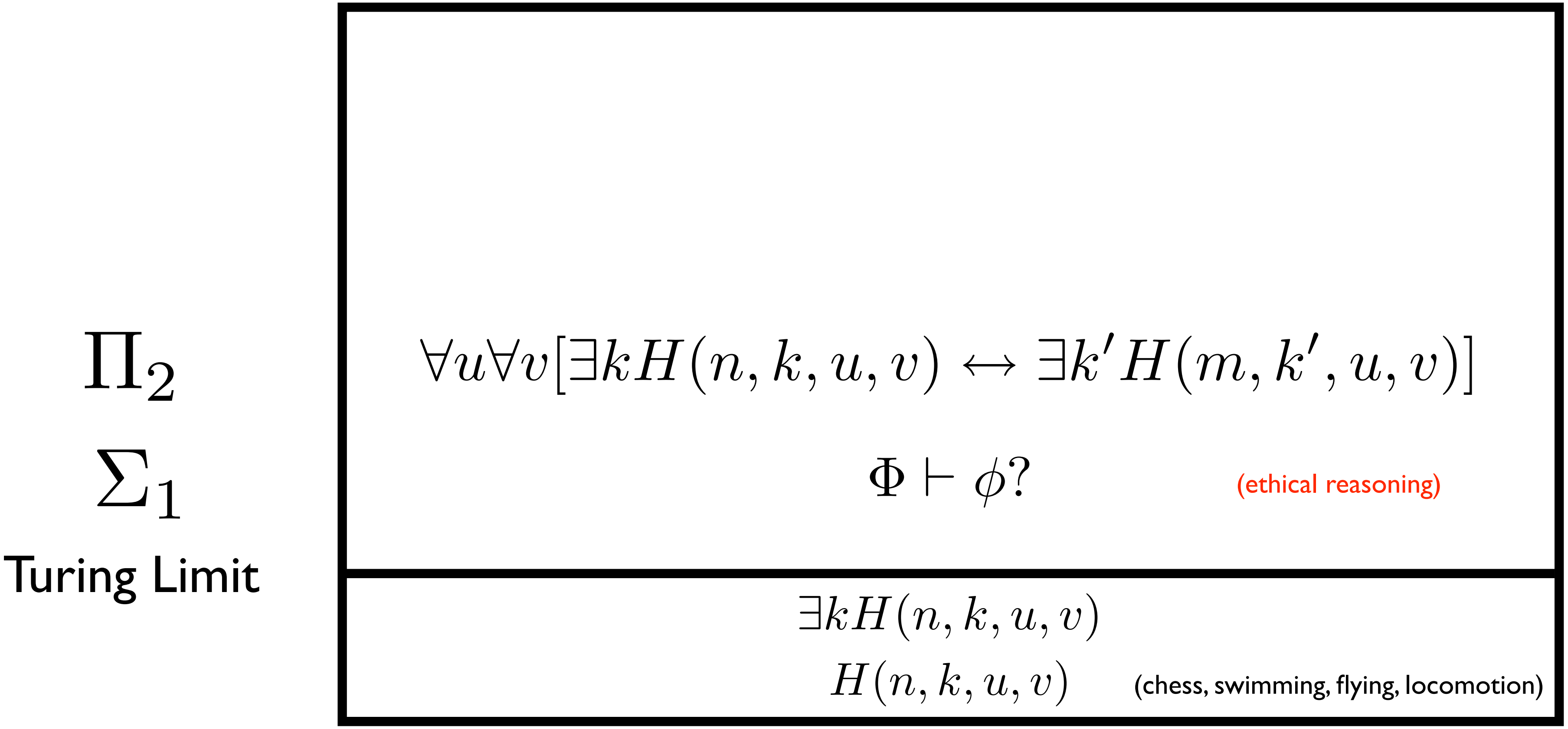
$$H(n, k, u, v)$$

(chess, swimming, flying, locomotion)

Beyond Reach of Turing Machines

$$\{f|f : N \to N\}$$

(Information Processing)



Π_2

Σ_1

Turing Limit

$$\forall u \forall v [\exists k H(n,k,u,v) \leftrightarrow \exists k' H(m,k',u,v)]$$

$$\Phi \vdash \phi?$$

(ethical reasoning)

$$\exists k H(n,k,u,v)$$

$$H(n,k,u,v)$$

(chess, swimming, flying, locomotion)

New Question

What could possibly be an alternative approach to solving the problem?

Logic is Our Only Hope

We only have one way to fix the meaning of programs, to verify that they will behave as advertised.

Enumerative induction will get us killed.

Logic is our only hope, ladies and gentlemen.

Logic is Our Only Hope

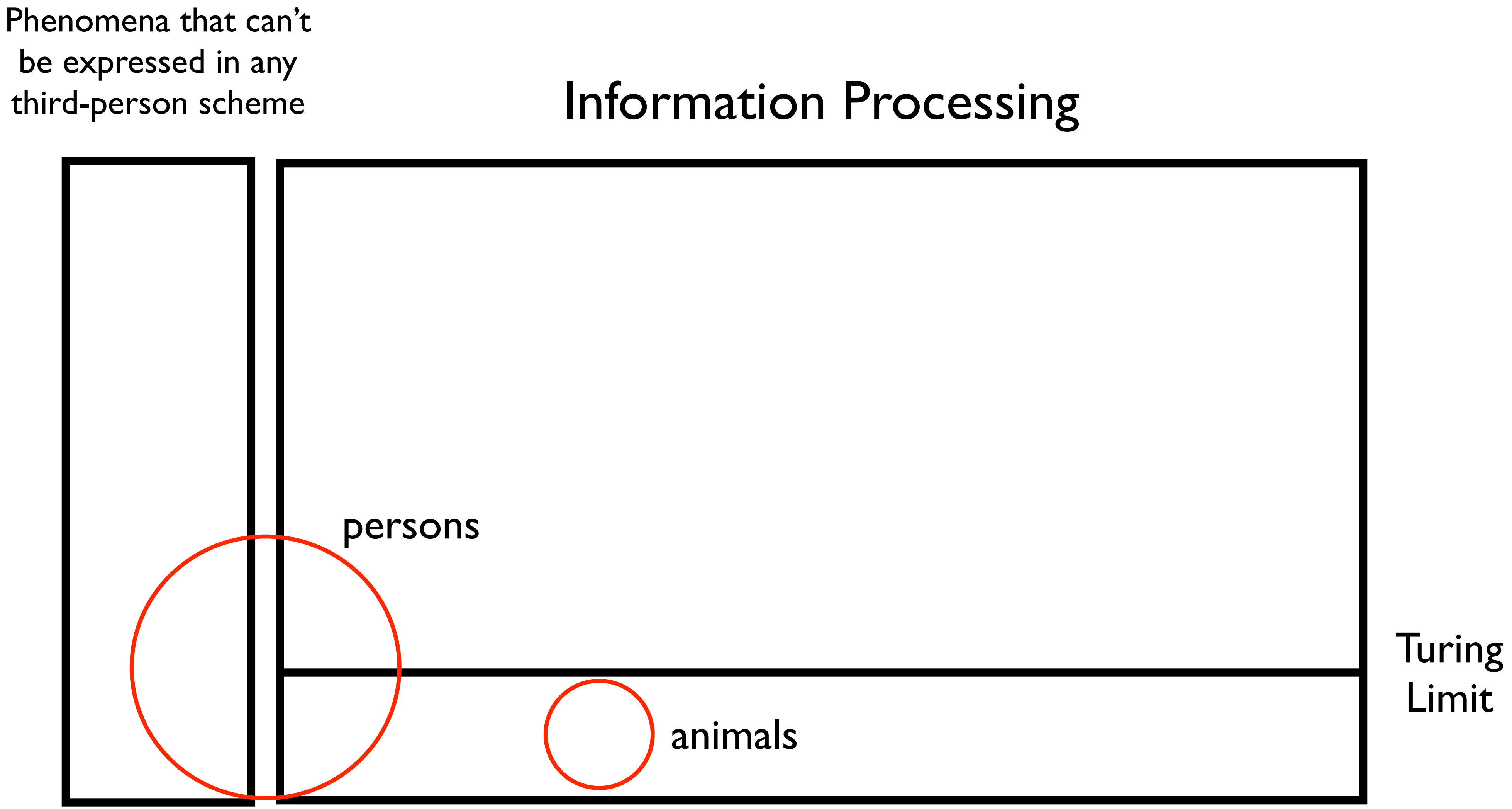
We only have one way to fix the meaning of programs, to verify that they will behave as advertised.

Enumerative induction will get us killed.

Logic is our only hope, ladies and gentlemen.

Finis

Superminds (2003)



Phenomena that can’t be expressed in any third-person scheme

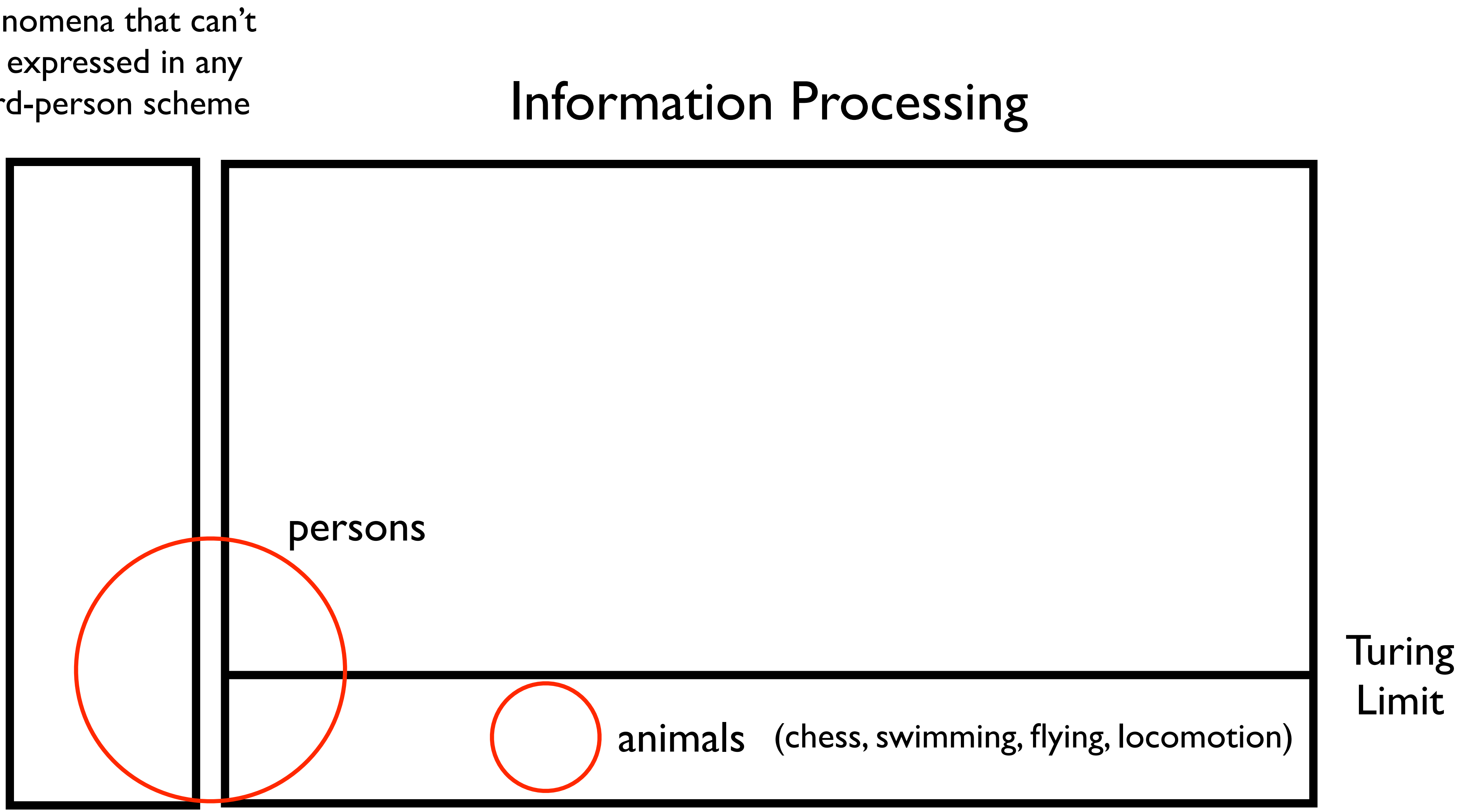
Information Processing

persons

animals

Turing Limit

Superminds (2003)



Phenomena that can't be expressed in any third-person scheme

Information Processing

persons

Turing Limit

animals (chess, swimming, flying, locomotion)