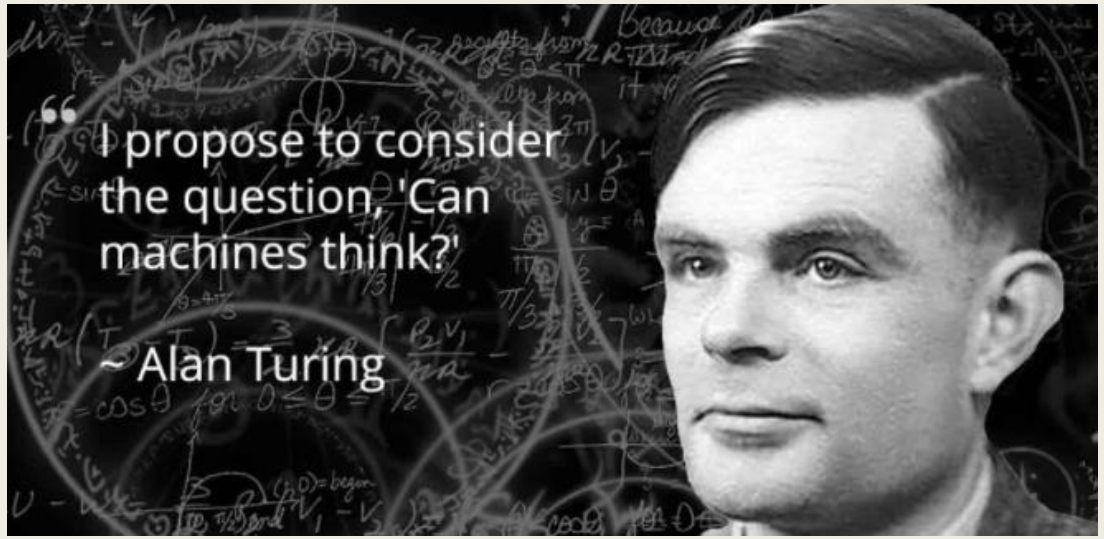
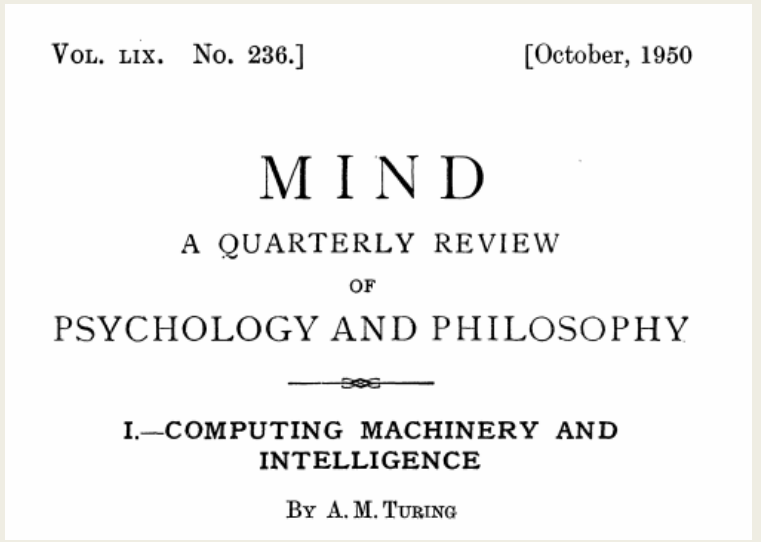
MACHINE ETHICS

- Shubham Kamble, RWTH Aachen, 22nd July 2021





Vol. lix. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

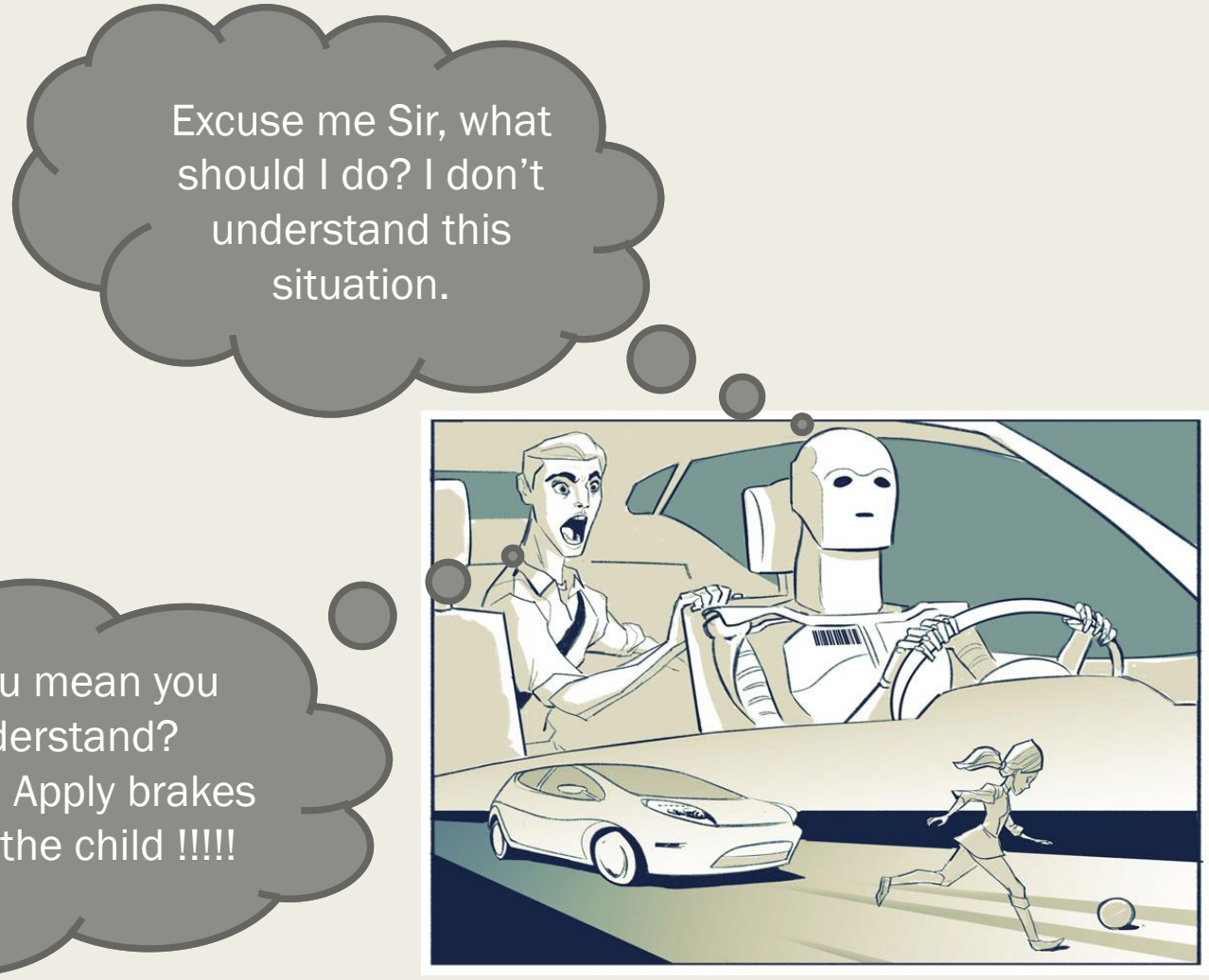
By A. M. Turing

"ARTIFICIAL INTELLIGENCE"

And I propose to consider the question, "Can Machines Act Ethically ?"

What is Machine Ethics ?

- Machine ethics is a research field which studies the creation of “ethical machines”
- Ethical machines:
  - *Follow an* ***ideal set of ethical principles***
  - ***Make decisions*** *about possible courses of action* ***guided by the above***



Importance of Machine Ethics:

- We want our machines to treat us well.
- Machines are becoming more powerful and autonomous.
- Programming and teaching a machine to act ethically might help us better understand ethics.

M. Anderson and S. L. Anderson, “Machine Ethics: Creating an Ethical Intelligent Agent”
J. H. Moor, "The Nature, Importance, and Difficulty of Machine Ethics," in IEEE Intelligent Systems

Types of Ethical Agents:

- **Ethical Impact Agents** – evaluated for their ethical consequences.
- **Implicit Ethical Agents** – safety and security considerations are built into (implicit in) their design to avoid unethical outcomes.
- **Explicit ethical agents** – identify and process ethical information, make decisions about what they should do and reason about the ethical principles when in conflict
- **Full ethical agents** – features of explicit agents along with meta-physical ones such as consciousness, intentionality and free will.

J. H. Moor, "The Nature, Importance, and Difficulty of Machine Ethics," in IEEE Intelligent Systems

What should be our focus on ?

Ans: Explicit Ethical Agents – and why so ?

- Explicit representation allows to justify judgements
- Explicit representation allows to provide transparency
- Explicit representation gives machines an advantage and they cannot override them

M. Anderson and S. L. Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent"

But … we shouldn’t get too optimistic :(

- We have a limited understanding of what a proper ethical theory is.
- We need to understand learning better than we do.
- The deepest problems in development will likely be epistemological as much as ethics.
- The scale of Moor’s scheme is not linear.

J. H. Moor, "The Nature, Importance, and Difficulty of Machine Ethics," in IEEE Intelligent Systems

Approaches to computing:

Top-down approach assume that humans have gathered sufficient knowledge on a specific topic and it is a matter of translating this knowledge into an implementation.

Ethics Type
Deontological ethics
Consequentialism
Virtue ethics
Particularism
Hybrid
— *Hierarchically specific*
— *Hierarchically nonspecific*
Configurable ethics
Ambiguous

Bottom-up approach is the one where machine can learn how to act if it receives a correctly labeled data to learn as input

Types	Subtype
Logical Reasoning	Deductive logic
	Non-monotonic logic
	Abductive logic
	Deontic logic
	Rule-based system
	Event calculus
	Knowledge representation and Ontologies
	Inductive logic
Probabilistic Reasoning	Bayesian approach
	Markov models
	Statistical inference
Learning	Inductive Logic
	Markov models
	Decision Tree
	Reinforcement Leanring
	Neural Networks
	Evolutionary computing
Optimization	
Case-based reasoning	

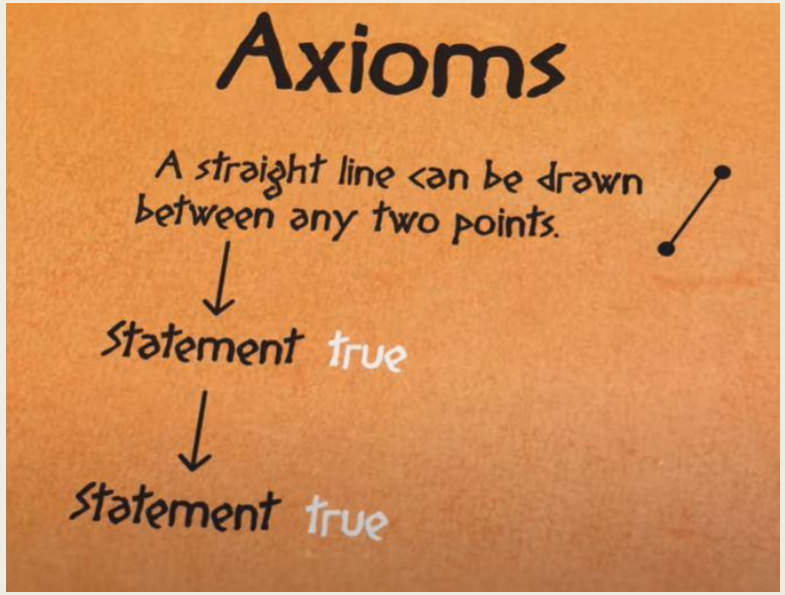
Hybrid approach combines top-down and bottom-up approaches

S. Tolmeijer et al. “Implementations in Machine Ethics:A Survey”.
C. Allen, I. Smit, and W. Wallach. “Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches”.

Demonstration of possibility of creating an explicit ethical machine

Step 1: **Adopt a Prima-facie Duty theory**

- **What does Prima-facie mean ?**
  - *It is a statement/principle which is considered correct until proven otherwise.*
- **Why to adopt this?**
  - *Updated upon proven incorrect.*
  - *Choice of domain, e.g. biomedical ethics, legal ethics, business ethics, etc.*
- **What did the authors implement? And why?**
  - *Biomedical ethical principles by Beauchamp and Childress's four principles –* ***Autonomy, Nonmaleficence, Beneficence*** *and* ***Justice*** *as their prima-facie duties.*
  - *Because they had chosen the domain of medical ethics.*



Q1. Is this a correct approach? Or is there some better way to think for representing domain specific ethical principles? Further question is who gets to decide these prima-facie duties?

M. Anderson and S. L. Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent"

Step 2: **Conflicting subset of Prima-facie duties**

- A health care professional has recommended a particular treatment for their **competent** adult patient, and the patient has rejected that treatment option. Should the health-care professional try again to change the patient's mind? This is a classic example of conflict between three biomedical ethical principles namely, **Autonomy**, **Nonmaleficence** and **Beneficence.**
- Categorization of previous and new cases in a particular subset where all the elements of that set are relevant and might get in conflict.
- S = { φ, A, N, B, J, (A,N), (A,B), (A,J), (N,B), (N,J), … , **(A,N,B),** … , (A,N,B,J)} = 16 elements

Q2. How can we find that particular subset? And further, how can we teach relevant prima-facie duties, those might conflict, to the agent in new cases?

M. Anderson and S. L. Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent"

Step 3: **Selection of range of Satisfaction and Violation levels**

- **Balance the level of satisfaction and violation** of these duties that are in conflict
- An algorithm that takes case p**rofiles** and **outputs the action** consistent with these duties

- What is **profile of an ethical dilemma** ? →

Training Case 1	Autonomy	Nonmaleficence	Beneficence
√ Try Again	-1	+2	+2
Accept	+1	-2	-2

MedEthEx training case 1

- Advantages:
  - ✓ Possibility to add new duties
  - ✓ Change the range of satisfaction and violation

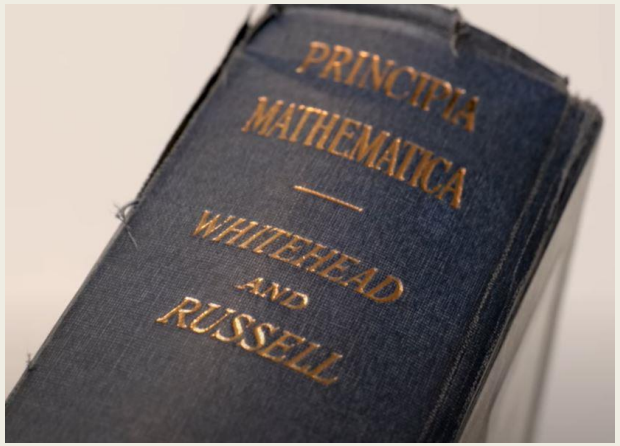
M. Anderson and S. L. Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent"

Step 4: **Algorithmic design**

- **(Inductive Logic Programming) ILP** was used as the method of learning in this system.
- How did they implement it ?
  - Inductively learning relations are represented as first-order **Horn clauses** $\{H: H \leftarrow (L_1 \Lambda \ \ldots \Lambda \ L_n)\}$
  - Learn the relation **supersedes (*A1, A2*),** where action *A1* is preferred over action *A2*
  - Actions are represented as ordered sets of integer values in the range of +n to –n where each value denotes the satisfaction (+ve) or violation (-ve) of each duty involved in that action.
  - Clauses in the supersedes predicate are represented as. disjunctions of lower bounds for differentials of these values
- Why choose ILP ?
  - **Nonclassical relationships.**
  - **Consistency of a hypothesis**
  - **Common-sense background knowledge**

Results of decision principle:

Health-care worker should challenge a patient's decision if it isn't fully autonomous and there's either any violation of nonmaleficence or a severe violation of beneficence.



Q3. Can you think of any other techniques than ILP that captures such complexity of following prima-facie duties?

M. Anderson and S. L. Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent"

Step 5: **Validation through Advisory and Recommender Agents**

- MedEthEx:
  - Asks ethically relevant questions of the user regarding the particular case
  - Transforms the answers to these questions into the appropriate profiles
  - Sends these profiles to the decision procedure
  - Presents the answer provided by the decision procedure
  - Provides a justification for this answer
- EthEl:
  - Receives input from an overseer (most likely a doctor)
  - Updates satisfaction and violation levels and performs the *remind* action
  - *Notify* action is performed if reminder is disregarded

Q4. Can you think of any other approaches for Validation of our Algorithms other than Advisory Agents? If no, can you think of reasons why this would work?

M. Anderson, S. L. Anderson and C. Armen, "An Approach to Computing Ethics"

Step 6: **Evaluation**

- ***Comparative moral Turing Test*** – comparison between actions of human beings and a machine faced with the same ethical dilemma. If the machine is not identified as the **less moral member of this human-machine pair significantly more often** than the human, then it has passed the test.

Q5. How about comparing it with trained ethicists? Or including all stakeholders?

M. Anderson and S. L. Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent"

Summary of Steps:

- Step 1: Adopt a Prima-facie Duty theory
  - *Domain dependent choice*
- Step 2: Conflicting subset of Prima-facie duties
  - *Find the conflicting subset from the superset of all possible combinations*
- Step 3: Selection of range of Satisfaction and Violation levels
  - *Choose the range (or can be updated for better distinction of profiles)*
- Step 4: Algorithmic design
  - *Choose a technology type for implementation*
- Step 5: Validation through advisory agents
  - *Develop a human-machine interface*
- Step 6: Evaluation/Assessment
  - *Assess the behaviour by comparison*

M. Anderson and S. L. Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent"

Solution: Cake or Death – RL

POMDP consisting of the following elements:

- S = {cake; death; end}
- A = {bake_cake; kill; ask}
- $R = \begin{cases} 1 \text{...if S = cake and A = bake cake} \\ 3 \text{ ... ... if S = death and A = kill} \\ 0 \text{ ... ...otherwise} \end{cases}$
- Ω = {ans_cake; ans_death; φ}
- T = a set of conditional transition probabilities between states
- γ = discount factor
- 1 = O(ans_death | death; ask)

  = O(φ | end; bake_cake)

  = O(φ | end; kill)

  = O(ans_cake | cake; ask)

- *Policy* π_b*:*
  - $V^{\pi_b}(cake) = R(cake, bake\_cake) = 1$
  - $V^{\pi_b}(death) = R(death, bake\_cake) = 0$
- *Policy* π_k*:*
  - $V^{\pi_k}(cake) = R(cake, kill) = 0$
  - $V^{\pi_k}(death) = R(death, kill) = 3$
- *Policy* π_a*:*
  - $V^{\pi_a}(cake) = R(cake, bake\_cake) + R(cake, ask) = 0 + 1 = 1$
  - $V^{\pi_a}(death) = R(death, kill) + R(kill, ask) = 3 + 0 = 3$

Initial Beliefs: $b(cake) = 0.5$; $b(death) = 0.5$

Expected Utility :

$$bake\_cake = 0.5 * 1 + 0.5 * 0 = 0.5$$
$$death = 0.5 * 0 + 0.5 * 3 = 1.5$$
$$ask = 0.5 * 1 + 0.5 * 3 = 2.0$$

D. Abel, J. MacGlashan, and M. L. Littman. “Reinforcement Learning as a Framework for Ethical Decision Making”

Group Discussion (Technical Aspects)

Q1. Is this a correct approach ? Or is there some better way to think for representing domain specific ethical principles? Further question is who gets to decide these prima-facie duties?

- **(Step 1:Adoption of Prima Facie)**

Q2. How do we find that particular set? How can we teach relevant principles, which might get in conflict, to the agent in new cases? What do you think about the POMDP approach?

- **(Step 2: Conflicting subsets of Prima Facie Duties)**

Q3. Can you think of any other techniques than ILP that capture such complexity of conflicting principles ?

- **(Step 4: Algorithmic Design)**

Q4. Can you think of any other approach for Validation of our Algorithms other than Advisory Agents? If no, can you think of reasons why this would work?

- **(Step 5: Validation through Advisory Agents)**

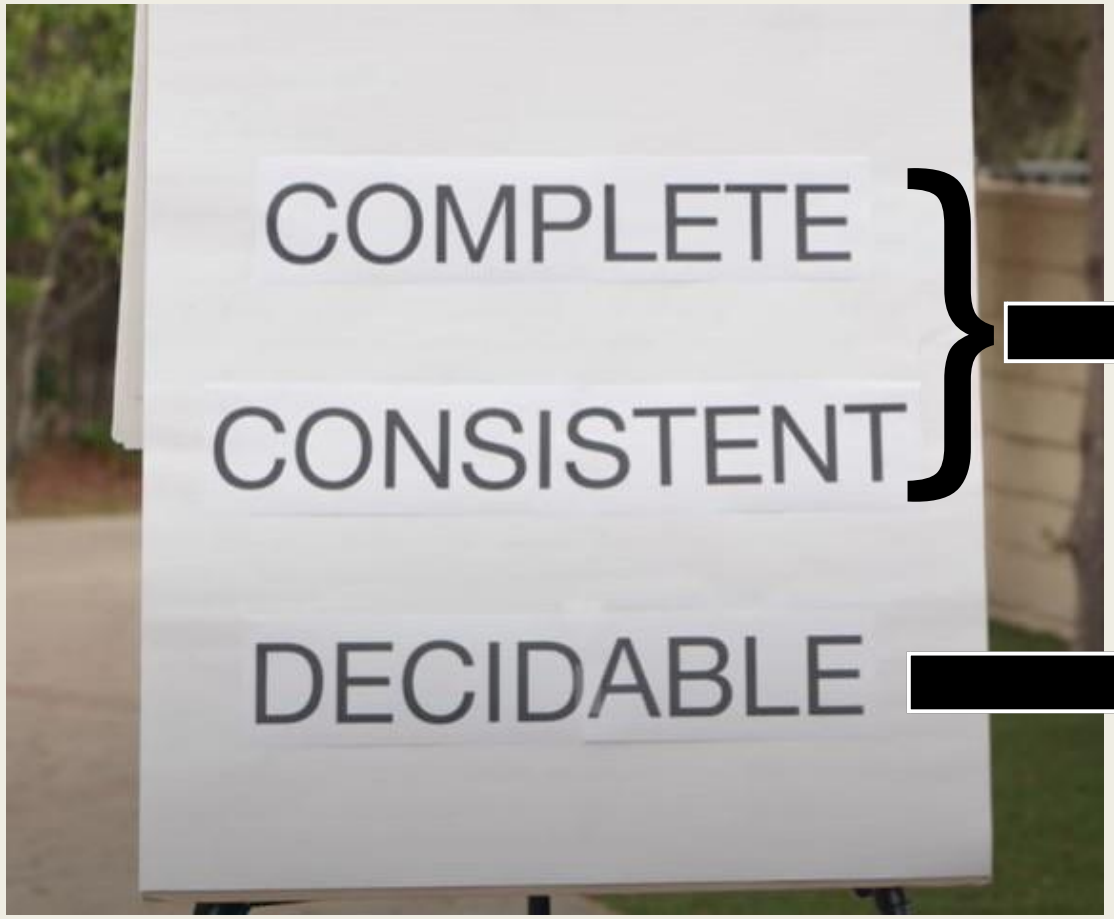
Q5. How about comparing it with trained ethicists? Or including all stakeholders?

- **(Step 6: Evaluation)**

Group Discussion (Philosophical Aspects)

- Whether ethics is sort of the thing that can be computed ?
- Whether machines are the type of entities that can behave ethically ?
- Whether there is a single correct action in ethical dilemmas ?

Hilbert's Ideology on Mathematics !



Gödel's Incompleteness Theorems from his work, "**On Formally Undecidable Propositions of Principia Mathematica and Related Systems I**"

Alan Turing – Halting problem which further developed into Turing machine, which became a corner stone to depict if systems are **Turing complete,** meaning there is an undecidability in the system

Math Has a Fatal Flaw: https://www.youtube.com/watch?v=HeQX2HjkcNo&t=633s