

חלק א- הצגת המאמר (עד 45 דקות)

1. דברי פתיחה- כמה משפטים כלליים על המאמר ועל המחבר, הצהרת כוונות שלי לגבי ההרצאה והצגת הנושאים העיקריים שעליהם נעבור (סה"כ 2-3 דקות)  
חלקים 1-4 של המאמר: הגדרת המושגים שבהם נשתמש בהמשך כשהמטרה היא להציג את המונחים וההגדרות כפי שהופיעו במאמר המקורי. נגדיר סמלים מסוג אחד וסמלים מסוג שני, מספרים חשיבים וסדרות חשיבות מכונות circular free-i circular choice/automatic (סה"כ 5 דקות)
2. חלק 5- נגדיר S.D | Standard Description – D.N | Description Number – D.N. נראה דוגמא- נציג מכונה פשוטה. נראה כיצד ניתן לקבל את ה-S.D מתוך הטבלה שלה ומתוכו להוציא את ה-D.N. נוכיח כי הסדרות והמספרים החשיבים הינם בני מניה דבר שימש אותנו בהמשך להוכחת אי-כריעות בעיית העצירה (סה"כ כ-10 דקות תלוי ברמת הפירוט שאגיע אליה בתיאור המכונה והקידוד).
3. חלקים 6-7- תיאור כללי של המכונה האוניברסלית. מה היא עושה ולמה זה טוב (2 דקות)
4. חלק 8- (סה"כ 15 דקות)
  - a. ראשית נביא את הוכחת אי-כריעות בעיית העצירה בצורה המוכרת שלה- "הוכחה בשתי דקות" (2 דקות).
  - b. אח"כ אראה את ההוכחה כפי שהיא מופיעה במאמר (כ-8 דקות).
  - c. לבסוף נראה תוצאה של b- לא קיימת מכונה שבהינתן S.D של מכונה אחרת יכולה לומר האם היא אי פעם סמל כלשהו- נניח 0 (3 דקות).
5. חלקים 9-10- (כ-10 דקות)
  - a. נדבר על הקונספט של חישוב כפי שטיורינג מציג אותו במאמר. נראה קצת נימוקים אינטואיטיביים לתיזה של טיורינג-צ'רץ' (2-3 דקות).
  - b. נראה הגדרה שקולה למושג של חישוב ע"י נוסחאות מנוסחות היטב (W.F.F) (3 דקות)
  - c. נדבר קצת על קבוצות של מספרים חשיבים. יהיו עוד (קצת) הגדרות Computable functions & Computable Convergence (2-3 דקות)
  - d. נראה דוגמא לפונקציה לא חשיבה בשלמים – דוגמא לא פורמלית (2-3 דקות)
6. חלק 11- "הפואנטה" של המאמר נציג את התוצאה הסופית כנובעת מחלק 8 (של המאמר). נדבר על היחסים בין בעיית העצירה לבעיית ההכרעה למשפטי אי השלמות של גדל. (3 דקות).
7. נספח- כמה מלים על המודל של צ'רץ' (effective calculability) ועל השקילות למודל של טיורינג בלי להיכנס להוכחה המדוייקת (דקה 1)

חלק ב- דברים שמכונות טיורינג יכולות לעשות. דברים שאינן יכולות לעשות? (עד 35 דקות)

1. האם מכונות יכולות להתרבות? נראה מכונה שכותבת את התיאור של עצמה- משפט הריקורסיה של קליני (15 דקות).
2. האם מכונות יכולות לחשוב? "האם צוללת יכולה לשחות?" נדבר קצת על מבחן טיורינג.
3. הקונספט הכללי, קצת על תחרויות ותוצאות עדכניות (כ-5 דקות)  
נדבר קצת על קוגניציה- המדע של המיינד. נדבר על המודל של המוח כחומרה והמיינד כתוכנה. (2-3 דקות)
4. נציג את "Chinese Room Argument" בתור counter argument ל-3. נראה טיעונים שעלו כנגד טיעון החדר הסיני-"החדר מבין!" (10-5 דקות)

סה"כ זמן ההרצאה המתוכננת: כשעה ורבע (לא כולל זמן לדיון חופשי ושאלות- נניח עוד רבע שעה לכל היותר).

## ON COMPUTABLE NUMBERS, WITH AN APPLICATION TO THE ENTSCHEIDUNGSPROBLEM

“There is no sense in being precise when you don't know what you are talking about...”  
- von Neumann

אז איך מתחיל הסיפור שלנו...

The epitaph on his tombstone in Göttingen consists of the famous lines he spoke at the conclusion of his retirement address to the Society of German Scientists and Physicians in the fall of 1930. The words were given in response to the Latin maxim: "Ignoramus et ignorabimus" or "We do not know, we shall not know".<sup>[21]</sup>

*Wir müssen wissen.*

*Wir werden wissen.*

In English:

We must know.

We will know.

**The day before Hilbert pronounced these phrases** at the 1930 annual meeting of the Society of German Scientists and Physicians, **Kurt Gödel**—in a roundtable discussion during the Conference on Epistemology held jointly with the Society meetings—**tentatively announced the first expression of his incompleteness theorem.**<sup>[22]</sup>

The origin of the Entscheidungsproblem goes back to **Gottfried Leibniz**, who in the seventeenth century, after having constructed a successful mechanical calculating machine, dreamt of building a machine that could manipulate symbols in order to determine the truth values of mathematical statements.<sup>[2]</sup> He realized that the first step would have to be a clean formal language, and much of his subsequent work was directed towards that goal. In 1928, **David Hilbert** and **Wilhelm Ackermann** posed the question in the form outlined above.

In continuation of his "program", **Hilbert** posed three questions at an international conference in 1928, the third of which became known as "Hilbert's Entscheidungsproblem".<sup>[3]</sup> As late as 1930, he believed that there would be no such thing as an unsolvable problem.<sup>[4]</sup>

### Gödel's first incompleteness theorem ...

first appeared as "Theorem VI" in Gödel's 1931 paper *On Formally Undecidable Propositions in Principia Mathematica and Related Systems I*.

The formal theorem is written in highly technical language. The broadly accepted natural language statement of the theorem is:

Any effectively generated theory capable of expressing elementary arithmetic cannot be both consistent and complete. In particular, for any consistent, effectively generated formal

theory that proves certain basic arithmetic truths, there is an arithmetical statement that is true,<sup>[1]</sup> but not provable in the theory (Kleene 1967, p. 250).

### קצת על הקשר שבין בעיית העצירה לבעיית ההכרעה...

The incompleteness theorem is closely related to several results about undecidable sets in recursion theory.

Stephen Cole Kleene (1943) presented a proof of Gödel's incompleteness theorem using basic results of computability theory. One such result shows that the halting problem is undecidable: there is no computer program that can correctly determine, given a program  $P$  as input, whether  $P$  eventually halts when run with a particular given input. Kleene showed that the existence of a complete effective theory of arithmetic with certain consistency properties would force the halting problem to be decidable, a contradiction. This method of proof has also been presented by Shoenfield (1967, p. 132); Charlesworth (1980); and Hopcroft and Ullman (1979).

### - על הבעיה ה-10 של דויד הילברט כפי שהוצגה בשנת 1900...

**Hilbert's tenth problem** is the tenth on the list of Hilbert's problems of 1900. Its statement is as follows:

Given a Diophantine equation with any number of unknown quantities and with rational integral numerical coefficients: *To devise a process according to which it can be determined in a finite number of operations whether the equation is solvable in rational integers.*

A Diophantine equation is an equation of the form

$$p(x_1, x_2, \dots, x_n) = 0,$$

where  $p$  is a polynomial with integer coefficients. It took many years for the problem to be solved with a negative answer. Today, it is known that no such algorithm exists in the general case. This result is the combined work of Martin Davis, Yuri Matiyasevich, Hilary Putnam and Julia Robinson.<sup>[1]</sup>

### משפט הריקורסיה של קליני

- האם מכונות יכולות להתרבות (סיפסר 230)?
- מכונות שמתרבות או ליתר דיוק "משכפלות" את עצמן- משפט+הוכחה
- הכללה- הצגת משפט הריקורסיה המוכלל + הוכחה אם יהיה זמן

## מבחן טיורינג

- האם מכונות יכולות לחשוב? כמו לשאול האם צוללת יכולה לשחות (דייקסטרה)?
- תיאור המבחן
- The Turing test is a test of a machine's ability to exhibit intelligent behaviour. In [Alan Turing's](#) original illustrative example, a human judge engages in a natural language conversation with a human and a machine designed to generate performance indistinguishable from that of a human being. All participants are separated from one another. If the judge cannot reliably tell the machine from the human, the machine is said to have passed the test.
- קצת על קוגניציה- המדע של המינד (cognitive science)
- עלייתו ונפילתו של מודל של המוח כחומרה והמינד כתוכנה. ביקורות על המודל...
- גישה דואליסטית. דומה בבסיסה למודל הראשוני של דקארט (מודל/בעיית גוף-נפש) בעצם לא מסבירה דבר.
- הטענה כי המוח אינו דטרמיניסטי
- מושג הבחירה- הרגשת חופש הבחירה ("רצון חופשי", החוויה האישית של כל אחד) אינה ניתנת להסבר באמצעות מכונת טיורינג דטרמיניסטית
- אפשר להגדיר מכונת טיורינג לא דטרמיניסטית, אבל ההגדרות שהוצעו אינן מדברות בכלל על מושג הבחירה. במאמר של טיורינג הבחירה היא שרירותית לחלוטין ונעשית בידי ערטילאית ולא מוגדרת.
- אפשר להגדיר מכונת טיורינג הסתברותית כפי שאכן עשו ועושים. יש לה חשיבות לא מבוטלת בסיבוכיות. בכל מקרה בין אם אנחנו בוחרים במודל של מכונה דטרמיניסטית או הסתברותית. אף אחד מהם לא מותיר מקום ל"רצון חופשי". אבל אפשר לומר שלפחות מבחינת המתבונן החיצוני, לגלגל את הקובייה פעם נוספת ולקבל תוצאה אחרת זה כמו לומר שאם היה אפשר לחזור בזמן ניתן היה לבחור בחירה שונה.
- ביקורת נוספת על המודל שנשמעת בחוגים של A.I. שמים מחשב בחדר חשוך בלי שום דרך לחוות את העולם. הוא אינו יכול לראות, אינו יכול לחוש ובודאי שאינו יכול להרגיש. אז איך מישהו יכול לצפות שיראה אינטליגינציה?
- טיעון החדר הסיני- פעולה מכניסטית שנותנת תשובות, מהירות ומדוייקות ככל שתהיינה, איננה בגדר חשיבה

If you can carry on an intelligent conversation with an unknown partner, does this imply your statements are understood?

Replies to Searle's argument may be classified according to what they claim to show:<sup>[1]</sup>

- Those which identify *who* speaks Chinese;
- Those which suggest that the Chinese room should be redesigned in some way;
- Those which contend that Searle's argument is misleading; and
- Those which argue that the argument makes false assumptions about subjective conscious experience and therefore proves nothing.

### System and virtual mind replies: finding the mind

These replies attempt to answer the question: since the man in the room doesn't speak Chinese, *where* is the "mind" that does? These replies address the key [ontological](#) issues of [mind](#)

[vs. body](#) and simulation vs. reality. All of the replies that identify the mind in the room are versions of "the system reply".

### **System reply**<sup>[49][a]</sup>

The basic "system reply" argues that it is the "whole system" which understands Chinese. While the man understands only English, when he is combined with the program, scratch paper, pencils and file cabinets, they form a system that can understand Chinese. "Here, understanding is not being ascribed to the mere individual; rather it is being ascribed to this whole system of which he is a part" Searle explains.<sup>[30]</sup> The fact that man does not understand Chinese is irrelevant, because it is only the system as a whole which matters.

### **Other minds and zombies: meaninglessness**[\[edit\]](#)

Several replies argue that Searle's argument is irrelevant because his assumptions about the mind and consciousness are faulty. Searle believes that human beings directly experience their consciousness, intentionality and the nature of the mind every day, and that this experience of consciousness is not open to question. He writes that we must "presuppose the reality and knowability of the mental."<sup>[90]</sup> These replies question whether Searle is justified in using his own experience of consciousness to determine that it is more than mechanical symbol processing. In particular, the other minds reply argues that we cannot use our experience of consciousness to answer questions about other minds (even the mind of a computer), and the epiphenomena reply argues that Searle's consciousness does not "exist" in the sense that Searle thinks it does.

### **Other minds reply**<sup>[91][a]</sup>

This reply points out that Searle's argument is a version of the [problem of other minds](#), applied to machines. There is no way we can determine if other people's subjective experience is the same as our own. We can only study their behavior (i.e., by giving them our own [Turing test](#)). Critics of Searle argue that he is holding the Chinese room to a higher standard than we would hold an ordinary person.

### **Speed and complexity: appeals to intuition**[\[edit\]](#)

The following arguments (and the intuitive interpretations of the arguments above) do not directly explain how a Chinese speaking mind could exist in Searle's room, or how the symbols he manipulates could become meaningful. However, by raising doubts about Searle's intuitions they support other positions, such as the system and robot replies. These arguments, if accepted, prevent Searle from claiming that his conclusion is obvious by undermining the intuitions that his certainty requires.

### **Speed and complexity replies**<sup>[83][a]</sup>

The speed at which human brains process information is (by some estimates) 100 billion operations per second.<sup>[85]</sup> Several critics point out that the man in the room would probably take millions of years to respond to a simple question, and would require "filing cabinets" of astronomical proportions. This brings the clarity of Searle's intuition into doubt.