

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



هوش مصنوعی پیشرفته

فصل ۲۶

# مبانی فلسفی هوش مصنوعی

**Artificial Intelligence: Philosophical Foundations**

کاظم فولادی  
دانشکده مهندسی برق و کامپیوتر  
دانشگاه تهران

<http://courses.fouladi.ir/aai>

# هوش مصنوعی

مبانی فلسفی هوش مصنوعی



## مقدمه

## فلسفه‌ی هوش مصنوعی

PHILOSOPHY OF ARTIFICIAL INTELLIGENCE

فیلسوفان، قرن‌ها پیش از اختراع کامپیوترها، برای حل مسائل هوش مصنوعی در تلاش بودند.

آیا ساختن یک هوش مصنوعی ممکن است؟

آیا ماشین می‌تواند مشابه افراد هوشمندانه کنش کند؟

آیا ماشین دارای ذهن است؟

پی‌آمد اخلاقی ماشین‌های هوشمند چیست؟

آیا پژوهش‌های هوش مصنوعی اخلاقی هستند؟

...

## فلسفه‌ی هوش مصنوعی

پرسش‌های بزرگ در مورد ذهن

### BIG QUESTIONS ABOUT MIND

ذهن‌ها چگونه می‌توانند کار کنند؟

ذهن‌های انسان‌ها چگونه کار می‌کنند؟

آیا ناهنسان‌ها ذهن دارند؟

## فلسفه‌ی هوش مصنوعی

### مسئله‌های اصلی

۲

#### هوش مصنوعی قوی

*Strong AI*

ادعای اینکه ماشین‌ها واقعاً فکر می‌کنند.

۱

#### هوش مصنوعی ضعیف

*Weak AI*

ادعای اینکه ماشین‌ها احتمالاً می‌توانند  
هوشمندانه کنش کنند.

اکثر پژوهشگران AI تا زمانی که فرضیه‌ی Weak AI پاسخ‌گوی برنامه‌هایشان باشد به فرضیه‌ی Strong AI توجه نمی‌کنند.

۳

#### پیامدهای اخلاقی

*Ethical Issues*

نسبت میان اخلاقیات و هوش مصنوعی

تمام پژوهشگران AI باید نگران پی‌آمدهای اخلاقی کارشان باشند.

# هوش مصنوعی

مبانی فلسفی هوش مصنوعی

۱

## هوش مصنوعی ضعیف:

آیا ماشین‌ها می‌توانند  
هوشمندانه کنش کنند؟

## هوش مصنوعی ضعیف

آیا ماشین‌ها می‌توانند هوشمندانه کنش کنند؟

### WEAK AI: CAN MACHINES ACT INTELLIGENTLY?

ماشین‌ها می‌توانند به گونه‌ای ساخته شوند تا  
به گونه‌ای کنش کنند که **گویی** هوشمند هستند؟؟

غیرممکن بودن هوش مصنوعی، وابسته به نوع تعریف آن است.

با تعریف:

« جستجو به دنبال بهترین برنامه‌ی عامل بر روی یک معماری داده شده »  
هوش مصنوعی ممکن است.

## آیا ماشین‌ها می‌توانند فکر کنند؟

### CAN MACHINES THINK?

### آیا ماشین‌ها می‌توانند فکر کنند؟

در مقایسه با پرسش‌های:

- \* آیا ماشین‌ها می‌توانند پرواز کنند؟
  - \* آیا ماشین‌ها می‌توانند شنا کنند؟
- این پرسش واضح طرح نشده است («تفکر»؟).

در عوض، تورینگ پیشنهاد دیگری داد:

آیا ماشین می‌تواند آزمون هوشمندی من را بگذراند؟  
(اینکه یک بازیس متبحر متقاعد شود که ماشین یک انسان است.)



## بحث از منظر «ناتوانی»

THE ARGUMENT FROM DISABILITY

ادعا: «یک ماشین هرگز نمی‌تواند  $X$  را انجام دهد»

واضح است که کامپیوترها خیلی از کارها را مانند انسان‌ها و یا بهتر از آنها انجام می‌دهند، اما کارهای زیادی هم وجود دارند که کامپیوترها هنوز در آن برتری ندارند.

MATHEMATICAL OBJECTION

پرسش‌های ریاضی خاصی هستند که  
اصولاً با سیستم‌های صوری خاص قابل پاسخ‌دهی نیستند.

## ایراد ریاضی

## مسئله‌ی توقف

HALTING PROBLEM

پرسش‌های ریاضی خاصی هستند که اصولاً با سیستم‌های صوری خاص قابل پاسخ‌دهی نیستند.

کامپیوترها نمی‌توانند مسئله‌ی توقف را حل کنند، اما انسان‌ها می‌توانند.  
(آیا برنامه‌ی  $P$  توقف می‌کند یا در حلقه‌ی نامتناهی گیر می‌افتد؟)

انسان‌ها می‌توانند از فرمالیسم فرار کنند و به پاسخ پرسش بپردازند.  
اما کامپیوترها نمی‌توانند (؟)

## ایراد ریاضی

قضیه‌ی ناکامل بودن گودل

GODEL'S INCOMPLETENESS THEOREM

پرسش‌های ریاضی خاصی هستند که اصولاً با سیستم‌های صوری خاص قابل پاسخ‌دهی نیستند.

با داشتن یک سیستم صوری  $F$   
 جمله‌ی  $G(F)$  (جمله‌ی گودل) وجود دارد که  
 (۱)  $G(F)$  سازگار با  $F$  است. (۲)  $F$  نمی‌تواند  $G(F)$  را ثابت کند.

انسان‌ها می‌توانند از فرمالیسم فرار کنند و  $G(F)$  را ثابت کنند.  
 اما کامپیوترها نمی‌توانند (؟)

یعنی: برخی از استدلال‌های انسان، الگوریتمیک نیست.

## ایراد ریاضی

پاسخ

MATHEMATICAL OBJECTION

پرسش‌های ریاضی خاصی هستند که اصولاً با سیستم‌های صوری خاص قابل پاسخ‌دهی نیستند.

۱) قضیه‌ی ناکامل بودن گودل فقط در سیستم‌های صوری با قدرت کافی برای حساب کاربرد دارد (مثل ماشین تورینگ / منطق مرتبه اول).

ماشین‌های تورینگ نامتناهی هستند اما کامپیوترها متناهی هستند و با یک سیستم بسیار بزرگ منطق گزاره‌ای قابل توصیف هستند، پس قضیه‌ی ناکامل بودن گودل آن را محدود نمی‌کند.

۲) یک عامل نباید از عدم توانایی در اثبات برخی جملات شرمسار باشد، در حالی که سایر عامل‌ها می‌توانند.

۳) حتی اگر تأیید شود کامپیوترها برای اثبات محدودیت دارند، دلیلی وجود ندارد که انسان‌ها برای اثبات محدودیت نداشته باشند.

## بحث از منظر «صوری نبودن»

### THE ARGUMENT FROM INFORMALITY

**ادعا:** «رفتار انسان بسیار پیچیده‌تر از آن است که بتواند در قالب قواعد بیان شود.»

اما کامپیوترها اساساً مبتنی بر قاعده هستند و  
کاری غیر از پیروی از قواعد نمی‌توانند انجام دهند،  
پس قادر به ایجاد رفتاری در حد هوشمندی انسان نیستند!

نظر فیلسوف هیوبرت دریفوس

اگر منظور، سیستم‌های مبتنی بر قاعده‌ی منطقی مرتبه‌ی اول بدون یادگیری باشد،  
حق با دریفوس است. در واقع این انتقاد متوجه یک روش خاص برنامه‌نویسی است!



*Hubert L. Dreyfus, 1929-*

# هوش مصنوعی

مبانی فلسفی هوش مصنوعی

## ۲

# هوش مصنوعی قوی:

آیا ماشین‌ها می‌توانند  
واقعاً فکر کنند؟



## هوش مصنوعی قوی

آیا ماشین‌ها می‌توانند واقعاً فکر کنند؟

### STRONG AI: CAN MACHINES REALLY THINK?

ماشین‌هایی که هوشمندانه کنش می‌کنند  
می‌توانند ذهن‌های واقعی خودآگاه داشته باشند؟؟

باورها، مطلوب‌ها و قصد‌ها، **درباره‌ی چیزی هستند**.

قصدمندی  
*Intentionality*

آگاهی عامل از حالت‌ها و کنش‌های ذهنی خود.

خودآگاهی  
*Consciousness*

## تز هوش مصنوعی قوی

### THE STRONG AI THESIS

#### تز هوش مصنوعی قوی (John Searle, 1980)

مجموعه‌ی همه‌ی ماشین‌های تورینگ ممکن (برنامه‌های کامپیوتری) را می‌توان از ساده‌تر تا پیچیده‌تر برشماری کرد:

$$T_1, T_2, \dots, T_k, \dots$$

اگر هوش مصنوعی ممکن باشد، حداقل یکی از این ماشین‌های تورینگ، یک هوش مصنوعی است (توجه داریم که پیچیدگی بی‌کران است ...)

#### هوش مصنوعی قوی:

در میان فهرست همه‌ی ماشین‌های تورینگ، همه‌ی هوش‌های ممکن وجود دارند:

$$T_1, T_2, \dots, T_{korb}, \dots, T_{einstein}, \dots, T_{andromedan}, \dots$$

## رویکردهای نظری در مورد حالتها و فرآیندهای ذهنی

## رویکردهای نظری در مورد حالتها و فرآیندهای ذهنی

۲

طبیعت‌گرایی زیست‌شناختی  
*Biological Naturalism*

حالت ذهنی، یک خاصیت سطح بالا ناشی از فرآیندهای سطح پایین نروبیولوژیکی است.

۱

نظریه‌ی کارکردگرایی  
*Functionalism Theory*

حالت ذهنی، هر وضعیت علّی واسط بین ورودی و خروجی است.

## مسئله‌ی ذهن-بدن

MIND-BODY PROBLEMمسئله‌ی ذهن-بدن  
*Mind-Body Problem*

نحوه‌ی ارتباط حالت‌ها و فرآیندهای ذهنی  
با حالت‌ها و فرآیندهای بدنی (به‌خصوص مغز) چگونه است؟

نسبت مسئله‌ی ذهن-بدن با رویکردهای نظری در مورد حالت‌ها و فرآیندهای ذهنی

۲

طبیعت‌گرایی زیست‌شناختی  
*Biological Naturalism*

فیزیکیالیسم، ماتریالیسم، مونئیسم، ...  
در مقابل دوآلیسم (دکارت)

شعار: مغزها باعث ذهن‌ها می‌شوند.

موانع نگاه ماتریالیستی:  
مسئله‌ی اختیار، موضوع خودآگاهی

۱

نظریه‌ی کارکردگرایی  
*Functionalism Theory*

تعریف مؤلفه‌ای مغز/ذهن:  
هر مؤلفه با توجه به خواص ورودی/خروجی  
آن تعریف می‌شود.

## مسئله‌ی ذهن-بدن

آزمایش مغز در خمره

BRAIN IN A VAT (BIAV) EXPERIMENT

## آزمایشی برای ارزیابی نظریه‌ی تناظر

- باورها، ساختارهایی در ذهن هستند.
- باورها به شاهد‌های حسگری ground شده‌اند.

نظریه‌ی تناظر  
*Correspondence Theory*

آیا می‌توان حالت‌های ارادی را از اشیای خارجی جدا کرد؟

*Intentional states*

مثل باور داشتن، دانستن، ترسیدن و ...

- مغز را از بدن جدا می‌کنیم و در یک خمره قرار می‌دهیم.
- ورودی و خروجی مغز را به کامپیوتر وصل می‌کنیم.
- کامپیوتر دنیای خیالی شبیه‌سازی شده را تولید و نگهداری می‌کند.
- اگر این آزمایش کار کند، ضرورتی برای وجود دنیای واقعی وجود ندارد!

⇐ ارزیابی روی‌کرد فیزیکالیسم

## مسئله‌ی ذهن-بدن

## آزمایش پروتز مغزی

BRAIN PROSTHESIS EXPERIMENT

## آزمایشی برای ارزیابی نظریه‌ی کارکردگرایی

آیا حالت‌های ذهنی در نرون‌های مغزی شکل می‌گیرند؟

- نرون‌های مغزی را به تدریج با همتهای الکترونیکی آنها جایگزین می‌کنیم.
- همتهای رفتار ورودی / خروجی مشابه نرون‌های مغزی دارند.
- آیا خودآگاهی سوژه تغییر پیدا می‌کند؟

⇐ ارزیابی روی کرد فانکشنالیسم

جان سرل (روی کرد فیزیکیالیسم) معتقد است که خودآگاهی سوژه از دست می‌رود.  
موراوک (روی کرد فانکشنالیسم) معتقد است که خودآگاهی بدون تغییر باقی می‌ماند.

## آیا ماشین‌ها می‌توانند فکر کنند؟

CAN MACHINES THINK?

آیا ماشین‌ها می‌توانند فکر کنند؟

ممکن نیست

*No Way**Hubert Dreyfus  
John Searle*

ممکن است

*Way*

بله

*Yes**Alan Turing*

## آیا ماشین‌ها می‌توانند فکر کنند؟

پاسخ آلن تورینگ به ایراد «لیدی آدا لاولیس»

### CAN MACHINES THINK?

آیا ماشین‌ها می‌توانند فکر کنند؟

ممکن نیست <i>No Way</i>	ممکن است <i>Way</i>	بله <i>Yes</i>
<i>Hubert Dreyfus</i> <i>John Searle</i>		<i>Alan Turing</i>

ایراد *Lady Ada Lovelace* (اولین برنامه‌نویس):

موتور تحلیلی هیچ همتی برای آغاز چیزی ندارد

آن فقط کاری را می‌تواند انجام دهد که ما می‌دانیم چگونه باید به آن بگوییم که انجام دهد.

پاسخ آلن تورینگ:

موتور تحلیلی قابلیت این را دارد که برنامه‌ریزی شود تا **یاد بگیرد** [یادگیری ماشین].

بنابراین، کاملاً قادر است چیزهایی را انجام بدهد که ما نمی‌دانیم چگونه باید به آن بگوییم.



## آیا ماشین‌ها می‌توانند فکر کنند؟

پاسخ آلن تورینگ به ایراد ریاضی

### CAN MACHINES THINK?

آیا ماشین‌ها می‌توانند فکر کنند؟

ممکن نیست <i>No Way</i>	ممکن است <i>Way</i>	بله <i>Yes</i>
<i>Hubert Dreyfus</i> <i>John Searle</i>		<i>Alan Turing</i>

ایراد ریاضی:

گودل، تورینگ و دیگران نشان داده‌اند که هیچ سیستم صوری خالصی (مثل کامپیوتر) نمی‌تواند هر واقعیت ریاضی را ثابت کند (بداند) - شامل واقعیت‌هایی که انسان‌ها در مورد این سیستم‌ها می‌دانند - . بنابراین، انسان‌ها برتر از ماشین‌ها هستند.

پاسخ آلن تورینگ:

اگرچه مشخص شده است که محدودیت‌هایی برای قدرت هر ماشین خاص وجود دارد، اما این بدون هیچ‌گونه اثباتی بیان شده است که چنین محدودیت‌هایی به خرد انسانی اعمال نمی‌شود. اگر هوش مصنوعی را توسعه بدهیم، ممکن است واقعیت‌هایی را بفهمد که انسان‌ها نمی‌توانند.

## آیا ماشین‌ها می‌توانند فکر کنند؟

آزمون تورینگ

CAN MACHINES THINK?

آیا ماشین‌ها می‌توانند فکر کنند؟

ممکن نیست <i>No Way</i>	ممکن است <i>Way</i>	بله <i>Yes</i>
<i>Hubert Dreyfus</i> <i>John Searle</i>		<i>Alan Turing</i>

برای اجتناب از بحث بی‌پایان، تورینگ آزمون زیر را برای هوشمندی پیشنهاد کرد:

## آزمون تورینگ

*Turing Test*

## بازی تقلید

*The Imitation Game*

پشت یک پرده، یک انسان و پشت پرده‌ی دیگر یک کامپیوتر است.  
 بازپرس انسانی از طریق تله‌تایپ با این دو در ارتباط است.  
 سوالاتی را تایپ می‌کند و پاسخ‌ها را روی نمایشگر می‌خواند.  
 اگر بعد از ۵ دقیقه بازپرس نتواند بفهمد پاسخ از کدام طرف می‌آید، کامپیوتر هوشمند است.

## The Turing Test: ELIZA

Eliza: Hello. I am ELIZA. How can I help you?

You: I'd like to know if you are intelligent

Eliza: Do you wish that i'm intelligent?

You: Not especially.

Eliza: Please go on.

You: Well, I think you should go on,  
if you're intelligent enough.

Eliza: Do you wish that yours intelligent enough?

You: Goodbye!

Eliza: Please go on.

## The Turing Test: TIPS

– Loebner Prize winner, 1995

> Yo. I'm here.

– You say: I'd like to know if you are intelligent.

> I'm just sitting here talking to you.

> I don't got nothing better to do right now.

– You say: If you had something better to do, what would it be?

> What?

– You say: You don't get it?

> I don't know.

– You say: I guess you don't know much, eh?

> Huh?

> Maybe I just got too much on my mind.

> I got real trouble at work. I think I'm going

> to get fired.

– You say: I think I understand your boss's point of view.

> What?

– You say: Goodbye!

> Bye, now.

## آیا ماشین‌ها می‌توانند فکر کنند؟

آزمون تورینگ و شرایط لازم و کافی برای هوشمندی

### CAN MACHINES THINK?

آیا ماشین‌ها می‌توانند فکر کنند؟

ممکن نیست <i>No Way</i>	ممکن است <i>Way</i>	بله <i>Yes</i>
<i>Hubert Dreyfus</i> <i>John Searle</i>		<i>Alan Turing</i>

متأسفانه، آزمون تورینگ برای هوشمندی نه لازم است و نه کافی

### آزمون تورینگ

*Turing Test*

کافی نیست	لازم نیست
برنامه‌های بی کاربرد (مثل الیزا) به طور منظم به هوش انسان‌های زیرمعمولی نزدیک می‌شوند!	موجود می‌تواند هوشمند باشد، اما اصلاً از هیچ انسانی تقلید نکند!

## آیا ماشین‌ها می‌توانند فکر کنند؟

جایگزین‌های آزمون تورینگ

### CAN MACHINES THINK?

آیا ماشین‌ها می‌توانند فکر کنند؟

ممکن نیست <i>No Way</i>	ممکن است <i>Way</i>	بله <i>Yes</i>
<i>Hubert Dreyfus</i> <i>John Searle</i>		<i>Alan Turing</i>

آزمون‌های جایگزین برای آزمون تورینگ ارائه شده است:

<p><b>آزمون تورینگ تماماً کامل</b> <i>Totally Complete Turing Test</i></p> <p>نیازمند پردازش داخلی اطلاعات یکریخت با انسان، از همه نوع تا پایین‌ترین سطح اتمی</p>	<p><b>آزمون تورینگ تام‌تر</b> <i>Even More Total Turing Test</i></p> <p>نیازمند پردازش داخلی اطلاعات یکریخت با انسان</p>	<p><b>آزمون تورینگ تام</b> <i>Total Turing Test</i></p> <p>صفحه نمایش حذف و توانایی ربات برای تعامل با محیط مورد آزمون قرار می‌گیرد (بدن‌مندی).</p>
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------

ممکن است به برنامه‌هایی برسیم که به طور عادی بسیاری از انسان‌ها را گول می‌زنند!

## آیا ماشین‌ها می‌توانند فکر کنند؟

ایراد هیوبرت دریفوس

### CAN MACHINES THINK?

آیا ماشین‌ها می‌توانند فکر کنند؟

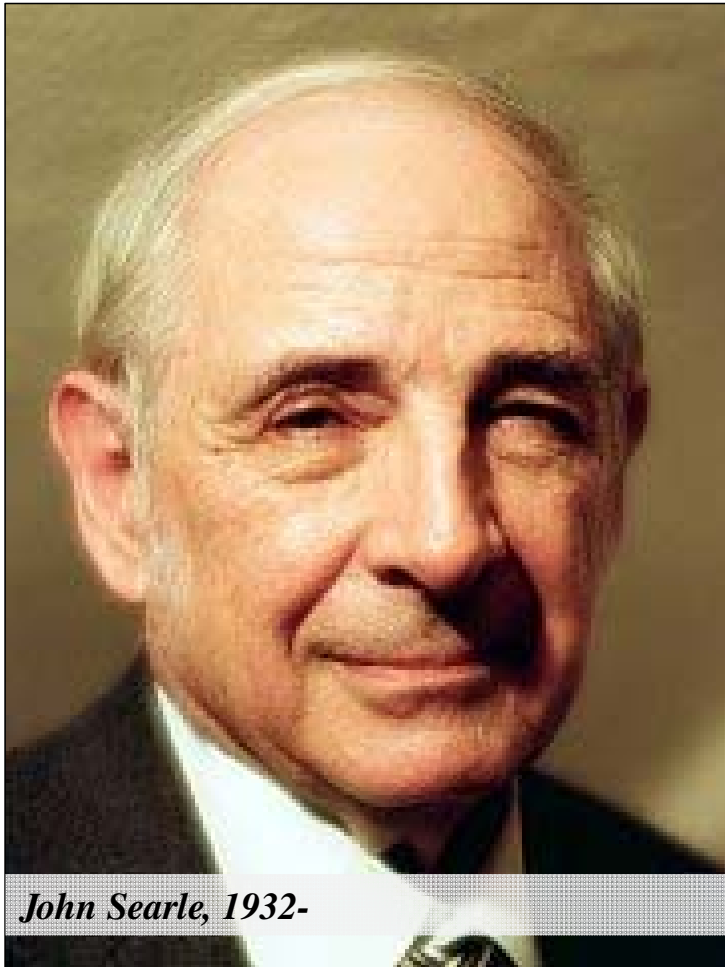
ممکن نیست <i>No Way</i>	ممکن است <i>Way</i>	بله <i>Yes</i>
<i>Hubert Dreyfus</i>		<i>Alan Turing</i>

ایراد *Hubert Dreyfus* (فیلسوف: 1972, 1990):

- خرد انسانی یک مهارت **فصل ناپذیر** (inarticulable) است؛ اما «خرد» کامپیوتر شامل **دنبال کردن قواعد** (rule-following) است.
- تفکر انسانی، دارای **زمینه** (context) است؛ اما قواعد کامپیوتر، **مستقل از زمینه** (context-free) است.
- تفکر انسانی **بدن مند** (embodied) است؛ اما اساس محاسبه عمومیت (یعنی **بدون جسم**: disembodiment) است.

- دنبال کردن قواعد تنها در سطح پیاده‌سازی مطرح است.
- لازم نیست قواعد یا رفتارها مستقل از زمینه باشند.
- ربات‌ها بدن مند هستند.

پاسخ



*John Searle, 1932-*



## آیا ماشین‌ها می‌توانند فکر کنند؟

## بحث اتاق چینی

CHINESE ROOM ARGUMENT

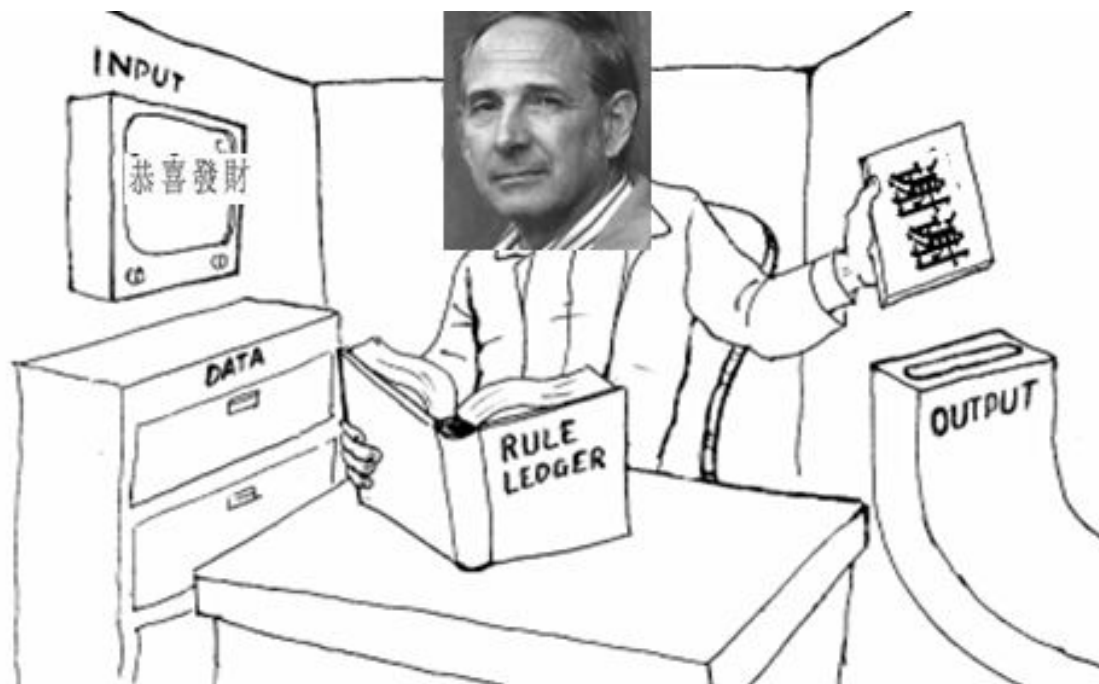
## آیا ماشین‌ها می‌توانند فکر کنند؟

ممکن نیست <i>No Way</i>	ممکن است <i>Way</i>	بله <i>Yes</i>
<i>John Searle</i>		<i>Alan Turing</i>

ایراد *John Searle* (فیلسوف: 1984) [آزمایش اتاق چینی]:

- فرض کنید یک برنامه‌ی کامپیوتری داریم که زبان چینی (مکتوب) را می‌فهمد.
- دستورات آن را در یک کتاب داخل اتاقی قرار می‌دهیم که فقط یک دریچه برای ورودی/خروجی مکتوب دارد.
- «جان سرل» [که چینی بلد نیست!] را داخل اتاق قرار می‌دهیم.
- هرگاه کاغذی وارد اتاق شد، «سرل» به کتاب نگاه می‌کند و کار مربوطه را مطابق کتاب انجام می‌دهد و کاغذ خروجی را بیرون می‌فرستد.

اتاق دارای رفتار زبانی مشابه یک مترجم چینی بومی است  
**اما هیچ فهمی در آنجا وجود ندارد!**



## آیا ماشین‌ها می‌توانند فکر کنند؟

پاسخ به بحث اتاق چینی

RESPONSE TO CHINESE ROOM ARGUMENT

آیا ماشین‌ها می‌توانند فکر کنند؟

ممکن نیست

No Way

John Searle

ممکن است

Way

بله

Yes

Alan Turing

پاسخ به ایراد *John Searle* (فیلسوف: 1984) [آزمایش اتاق چینی]:

## پاسخ خودآگاهی

*The Consciousness Response*

شهود سرل (نه خودآگاهی) بر اساس یک شهود عمیق‌تر و متفاوت‌تر است: هیچ خودآگاهی در آنجا نیست، اما هوشمندی بدون خودآگاهی یک امکان زنده است.

## پاسخ بدن‌مندی

*The Embodiment Response*

فهم نیازمند جاسازی علی است: ارتباط تله‌تایپ با دنیا کافی نیست. پاسخ اتاق چینی به این پرسش چیست؟ «رنگ لباس من چیست؟»

## پاسخ سیستمی

*The System Response*

هیچ فهمی داخل اتاق وجود ندارد، زیرا این اتاق است که به‌عنوان یک کل می‌فهمد!

# ۳

## اخلاقیات و مخاطرات توسعه‌ی هوش مصنوعی

## اخلاقیات و مخاطرات توسعه‌ی هوش مصنوعی

### THE ETHICS AND RISKS OF DEVELOPING ARTIFICIAL INTELLIGENCE

Russell & Norvig (2010):

*“If the effects of AI technology are more likely to be negative than positive, then it would be the moral responsibility of workers in the field to redirect their research.”*

اگر تأثیرات تکنولوژی هوش مصنوعی در بعد منفی، بیشتر از بعد مثبت آن است، وظیفه‌ی اخلاقی کسانی که در این زمینه کار می‌کنند، تغییر جهت پژوهش‌هایشان است.

## مشکلات احتمالی ناشی از گسترش هوش مصنوعی

ممکن است مردم به خاطر خودکارسازی ماشین‌ها کارشان را از دست بدهند.

ممکن است اوقات فراغت مردم خیلی کم (خیلی زیاد) شود.

ممکن است مردم احساس منحصر به فرد بودنشان را از دست بدهند.

ممکن است مردم برخی از حقوق حریم خصوصی‌شان را از دست بدهند.

ممکن است سیستم‌های هوش مصنوعی در جهت اهداف نامطلوب استفاده شوند.

ممکن است استفاده از سیستم‌های هوش مصنوعی منجر به فقدان مسئولیت‌پذیری باشد.

ممکن است موفقیت هوش مصنوعی به معنی پایان دوران نژاد انسان باشد.

Autonomous AI systems are now commonplace on the battlefield; the U.S. military deployed over 5,000 autonomous aircraft and 12,000 autonomous ground vehicles in Iraq (Singer, 2009).



*Joseph Weizenbaum, 1923-2008*



## آیا هوش مصنوعی اخلاقی است؟

دیدگاه وایزن باوم

### WEIZENBAUM'S ARGUMENT

- یک هوش مصنوعی واقعی در واقع باید یک عامل هوشمند خودمختار باشد (طبق تعریف)
- هیچ اشتراکی با ما نخواهد داشت: انگیزه‌ها، قیدها، اخلاقیات، ...
- هیچ کران بالایی واضحی برای هوشمندی نداریم؛ و احتمالاً اصلاً هیچ کران بالایی وجود ندارد.
- وقتی بین منافع ما و منافع هوش مصنوعی تضاد پیش می‌آید، حدس بزنید چه کسی می‌بازد؟

⇐ پژوهش در هوش مصنوعی غیراخلاقی است!

## ابر هوشمندی‌ها

دیدگاه وایزن باوم

SUPERINTELLIGENCES (SI)

هوش مصنوعی (بعدی) ساخته شده توسط هوش مصنوعی (اولیه)

تکینگی ابر هوشمندی

*SI Singularity*

می‌توانیم انتظار یک انفجار هوشمندی را داشته باشیم، وقتی که آستانه‌ی انسان به سر آید!

# FRANKENSTEIN ;

OR,

## THE MODERN PROMETHEUS.

---

IN THREE VOLUMES.

---

Did I request thee, Maker, from my clay  
To mould me man? Did I solicit thee  
From darkness to promote me?—

PARADISE LOST.

---

VOL. I.

---

London :

*PRINTED FOR*

LACKINGTON, HUGHES, HARDING, MAJOR, & JONES,  
FINSBURY SQUARE.

---

1818.

## قوانین رباتیک ایزاک آسیموف

### ASIMOV'S LAWS OF ROBOTICS, 1942

سه قانون رباتها عبارتند از:

۱ یک ربات نباید به نوع بشر صدمه بزند یا از طریق تنبلی موجب آسیب به نوع بشر شود.

۲ یک ربات باید از دستورهایی که توسط نوع بشر به او داده می‌شود، اطاعت کند.

۳ یک ربات باید وجود خودش را حفظ کند.

### فراقانون

ترتیب اولویت قانونها ۱، ۲، ۳ است.

### قانون صفرم

رباتها نباید به بشریت به‌عنوان یک کل آسیب وارد کنند.

داستان‌های آسیموف در مورد این است که چگونه این قوانین شکسته می‌شوند!

## ترنس اومانیزم

ماوراء بشرگرایی

TRANSHUMANISMترنس اومانیزم  
Transhumanism

ایدئولوژی جنبش اجتماعی فعالی که به دنبال آینده ماورای بشر است.

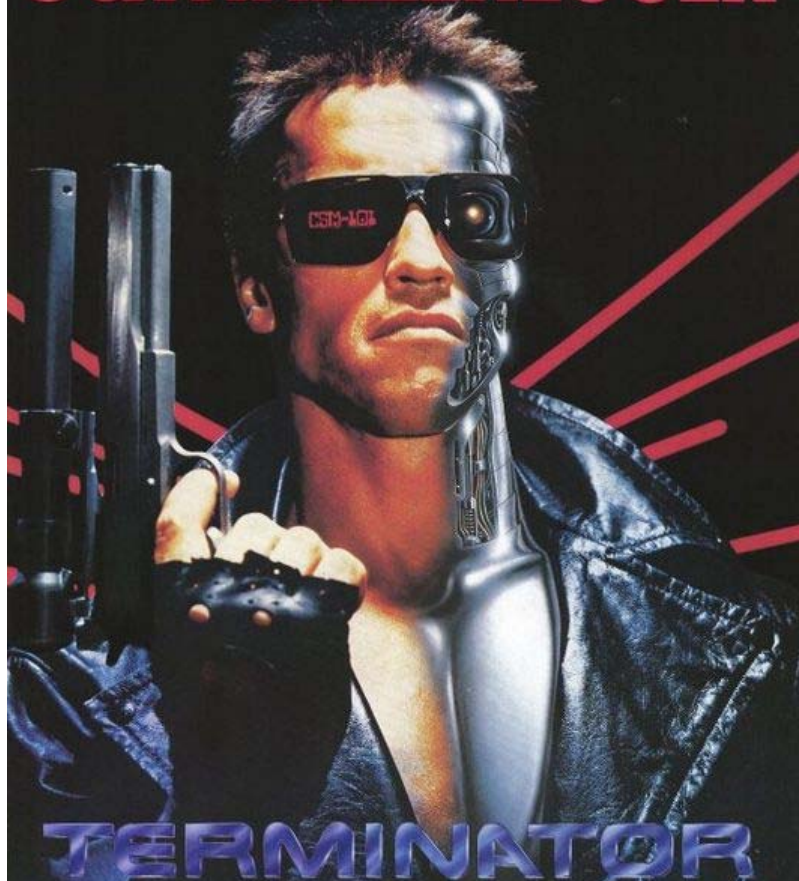
Ray Kurzweil, in *The Age of Spiritual Machines* (2000):

*By the year 2099 there will be “a strong trend toward a merger of human thinking with the world of machine intelligence that the human species initially created. There is no longer any clear distinction between humans and computers.”*

نزدیک سال ۲۰۹۹ یک روند قوی به سوی ادغام فکر بشری با دنیای هوش ماشینی که در ابتدا توسط نوع بشر پدید آمده است، وجود خواهد داشت. دیگر هیچ تمایز روشنی بین انسان‌ها و کامپیوترهای وجود ندارد.



# SCHWARZENEGGER



## TERMINATOR

Hemdale Presenta Una Produzione Pacific Western di un Film di James Cameron  
Arnold Schwarzenegger "Terminator" Michael Biehn, Linda Hamilton e Paul Winfield  
Effetti Speciali Trucco di Stan Winston · Direttore della Fotografia Adam Greenberg  
Produttore Esecutivo John Daly e Derek Gibson · Scritto da James Cameron con Gale Anne Hurd  
Prodotto da Gale Anne Hurd · Diretto da James Cameron

ORION

© 1984 Cinema '84. A Greenberg-Worhun Partnership. All rights reserved. Photo by [unreadable]

FREE YOUR MIND.

# THE MATRIX





HALEY JOEL OSMENT    JUDE LAW



A STEVEN SPIELBERG FILM

# A.I.

ARTIFICIAL INTELLIGENCE

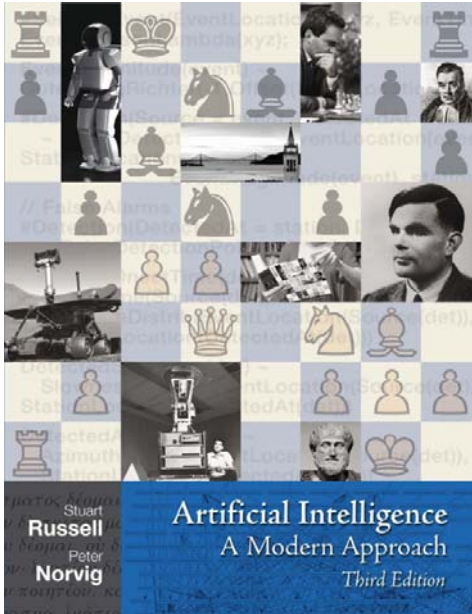


# هوش مصنوعی

مبانی فلسفی هوش مصنوعی

۴

منابع



Stuart Russell and Peter Norvig,  
**Artificial Intelligence: A Modern Approach,**  
 3<sup>rd</sup> Edition, Prentice Hall, 2010.

## Chapter 26

# 26 PHILOSOPHICAL FOUNDATIONS

*In which we consider what it means to think and whether artifacts could and should ever do so.*

Philosophers have been around far longer than computers and have been trying to resolve some questions that relate to AI: How do minds work? Is it possible for machines to act intelligently in the way that people do, and if they did, would they have real, conscious minds? What are the ethical implications of intelligent machines?

First, some terminology: the assertion that machines could act *as if* they were intelligent is called the **weak AI** hypothesis by philosophers, and the assertion that machines that do so are *actually* thinking (not just *simulating* thinking) is called the **strong AI** hypothesis.

Most AI researchers take the weak AI hypothesis for granted, and don't care about the strong AI hypothesis—as long as their program works, they don't care whether you call it a simulation of intelligence or real intelligence. All AI researchers should be concerned with the ethical implications of their work.

WEAK AI  
 STRONG AI

### 26.1 WEAK AI: CAN MACHINES ACT INTELLIGENTLY?

The proposal for the 1956 summer workshop that defined the field of Artificial Intelligence (McCarthy *et al.*, 1955) made the assertion that “Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it.” Thus, AI was founded on the assumption that weak AI is possible. Others have asserted that weak AI is impossible: “Artificial intelligence *pursued within the cult of computationalism* stands not even a ghost of a chance of producing durable results” (Sayre, 1993).

Clearly, whether AI is impossible depends on how it is defined. In Section 1.1, we defined AI as the quest for the best agent program on a given architecture. With this formulation, AI is by definition possible: for any digital architecture with  $k$  bits of program storage there are exactly  $2^k$  agent programs, and all we have to do to find the best one is enumerate and test them all. This might not be feasible for large  $k$ , but philosophers deal with the theoretical, not the practical.