# Conclusie Seminarie Turing Test

Sil Vaes

## Contents



## 1 Is de Turing Test nog relevant?

Niet echt, eerder een gedachtenexperiment. Ook raar dat hij een positivist was, maar de regels om te slagen zijn open voor interpretatie. Ook was schaken ook een van de intelligentietests, maar naar Deep Blue is dit geschrapt. Test niet echt intelligentie maar "menselijkheid". Er worden oppervlakkige conclusies getrokken. "Een computer kan gediagnosticeerd worden met koorts omdat de CPU een hoge temperatuur heeft". Ook raar dat een persoon met zijn overtuigingen, logisch positivisme, geen objectieve metriek heeft bedacht om de machine te beoordelen. Misschien een reden dat dit essay/paper gepubliceerd is in een blad voor filosofie.

Wat zijn dan wel moderne en relevante technieken om AI te evalueren? Twee soorten aanpakken:

- Task-oriented: met drie soorten evaluatie: human discrimination, problem benchmarks en peer confrontation.
- Ability-oriented: Waar een systeem beoordeeld word op zijn (cognitieve) vaardigheden ipv de taken die het kan vervolledigen. "the adaptation of cognitive tests used for humans and animals, the development of tests derived from algorithmic information theory or more integrated approaches under the perspective of universal psychometrics."

Waarom twee soorten aanpakken? Er zijn twee clashende definities van AI:

- "[AI is] the science and engineering of making intelligent machines" - McCarthy.
- "[AI is] the science of making machines capable of performing tasks that would require intelligence if done by [humans]" - Minsky.

De eerste is duidelijk ability-oriented en de tweede task-oriented. Iets ability-oriented evalueren is ook beduidend moeilijker. Bij task-oriented moet men gewoon kijken of de taak correct wordt uitgevoerd, lijkt ook meer op de Turing Test. Een AI voert een gespecialiseerde taak dan ook zonder intelligentie uit, wat een paradox eigenlijk. Dit is ook een reden van het succes, het is voorspelbaar. Zoals we in het begin van de cursus hebben gezien: "Iets is een AI probleem totdat het is opgelost". Dit wordt ook wel het "AI effect" genoemd (McCorduck). Voor dit seminarie is enkel het verschil tussen narrow en general AI belangrijk. Narrow of specialised AI heeft task-oriented evaluation nodig en general-purpose AI heeft ability-oriented evaluation nodig. Maar sommige AGI worden op specifieke taken getest en narrow AI's worden met general-purpose componenten zoals planning/learning geintegreerd en worden na trainingscycli gespecialiseerd en verliezen zo plasticiteit.

Zoals in alle wetenschap zijn metingen cruciaal. Momenteel zijn er benchmarks en competities, maar er is nog ruimte voor groei. Deels omdat men AI evaluatie niet echt als een meting ziet. Ook is het belangrijk om van de task-oriented mentaliteit af te geraken omdat er veel vooruitgang is in meer general AI met de opkomst van developmental robotics, deep learning, inductive programming, AGI, etc.

## 1.1 Task-oriented evaluation

Algemene principes worden toegepast, maar de taak is specifiek genoeg om een gespecialiseerd systeem te maken. Van een abstract probleem naar een specifieke taak wordt aangemoedigd. Wanneer we een specifieke taak hebben, moeten we een notie van performance hebben. Deze wordt dan ook gemeten. Het is duidelijk dat hier niet echt intelligentie wordt gemeten. Een AI voltooid deze taken ook op een heel andere manier dan mensen. We zeggen dan ook bijvoorbeeld dat het team achter Deep Blue Kasparov heeft verslaan.

Performance van een systeem wordt gemeten in een white-box approach als de oplossing relatief eenvoudig is. Een typisch probleem waarbij dit gedaan wordt is wanneer de oplossing correct of optimaal moet zijn (perfect dus). Dus eerder klassieke computational complexity theory.

Maar tegenwoordig zijn AI oplossingen zo complex dat perfecte oplossingen niet meer een constraint zijn. Dus zijn er approximate solvers die een performance metriek optimaliseren. Hier is average case analyse ook relevanter, maar bij sommige paradigma's wordt worst case ook bekeken (Probably Approximately Correct learning). White-box evaluation wordt ook vaak gebruikt bij spellen te spelen.

Maar tegenwoordig is AI zo complex geworden dat de white-box approach niet meer haalbaar is. Black-box approaches kunnen in drie categorien ondergebracht worden:

- Human discrimination: het AI wordt getest door observers. Komt niet vaak voor, hier valt de Turing Test onder. (Komt wel voor bij psychologie)
- Problem benchmarks: het systeem wordt getest op een repository van problemen. Dit komt vaak voor, er zijn libraries en bestaande repositories voor. Deze zijn ook vaak public, maar kan leiden tot "evaluation overfitting" (gespecialiseerd op deze problem-set).
- Peer confrontation: het systeem wordt getest in een serie van n-vs-n matches. Komt ook vaak voor, performance is relatief tov andere systemen. Vooral bij spellen en multi-agent systems.

Een combinatie van de drie kan ook. Common-sense reasoning kan door alle drie geevalueerd worden bijvoorbeeld. (interviewing, comprehension tests en een game show bijvoorbeeld).

**Human Discrimination**:

De Turing Test was bedoeld als een filosofisch instrument om tegen de contra argument van machine intelligence tegen te gaan, dus nooit bedoelt als een echte test. Het is een test die de menselijkheid test, niet intelligentie. Dit betekent niet dat we menselijke oordeel uit testen moeten halen. Er zijn varianten die wel nuttig zijn. (Total Turing tests (Schweizer 1998), Visual Turing tests including sensory information, Toddler Turing tests (Alvarado et al. 2002), robotic interfaces, virtual worlds, etc. (Mueller and Minnery 2008; Hingston 2010).)

- Truly Total Turing Test: Grootste argument tegen de TT is dat niet enkel het gedrag ertoe doet, maar hoe het gedrag gegenereerd wordt,

intelligentie is niet enkel performatief. Ook teveel nadruk op anonimiteit in de originele TT. Dus een Total Turing Test (TTT) waarbij het systeem de volledige menselijke activiteit kan naboots, dus er is een artificieel lichaam nodig. Maar dit is nog niet definitief, het kan nog nagebootst worden, kan enkel functioneren in een "toy world" door de designers van de test gemaakt. Dus nog niet echt "natuurlijke intelligentie". Zoals ons moet er een lang evolutionar criteria zijn. Wij zijn ook geevolueerd van rauwe omgevingsinputs naar rudimentaire sociale interactie en tool use naar agricultuur, kunst, wiskunde etc gegaan. Dit moet een AI ook kunnen. Dit is dan de Truly Total Turing Test (TTTT). Dit is niet voor individuele cognitieve systemen, maar voor om de capaciteit van een cognitieve architectuur. Dus een "samenleving", niet een token moet dit kunnen. Anders zouden zo goed als alle mensen ook de TTTT failen. Dus intelligentie is niet vraag van individuele samples, maar hoe dit gedrag wordt geproduceerd.

- Toddler Turing test: Deze test test of een systeem bepaalde essentiele, eerst vereiste cognitieve vaardigheden heeft. Hiervoor moet met men fundamentele vaardigheden identificeren die een systeem niet kan simuleren. Hoe wordt dit getest? Eerst, veranderingen in de interne parameters duiden op een verandering in mentale status. Ten tweede gebruikt met paradigma's uit de psychologie. Onobserveerbare cognitieve processen kan met afleiden uit kwantificeerbaar gedrag. Deze tests focussen zich op drie gebieden: associatief leren, sociale cognitie en taalverwerving.

- Turing Test voor Bots: Kan een echte AI bot spelers ervan overtuigen dat ze tegen een mens spelen? Het doel hiervan is dus daadwerkelijk om enkel de speler te overtuigen, hier is een echte Turing Test dus wel nuttig. Zie de BotPrize, waar een jury moet zeggen welke speler de mens was. Hier is het nuttig om evaluatiemethodes natuurlijker te maken door ze integreren in het spel zelf. (Men gebruikt unreal tournament 4) Er is een speciaal wapen dat enkel kan gebruikt worden als de speler/jurylid weet of hij tegen een bot of mens speelt. De speler krijgt punten wanneer hij een bot raakt en verliest wanneer het een mens is.

Deze zijn nuttig om chatbots, personal assistant, videogames etc te testen. Als in: zijn ze geloofwaardig?

**Problem benchmarks**:

Als de problemset klein en/of bekend is kan er een grote switch in het systeem gezet worden. Wanneer het systeem het probleem herkent kan het alsnog een hardwired oplossing gebruiken. -> Evaluation overfitting. Nog een groter probleem als de problemset gemaakt is door de researchers zelf. Bijvoorbeeld een self-driving car ontwikkelen op een kleine parking. Beter: een grote en diverse problemset. Nog beter: infinite, dus een generator. Maar hoe moeten we een infinite problemset evalueren? Eerste n problemen evalueren? Random sampling lijkt beter. Twee varianten:

- Information-driven sampling: Via clustering zo een divers mogelijk aantal problemen samplen via similarity.
- Difficulty-driven sampling: Moeilijke problemen samplen (via tijd nodig of error).

Beide manieren kunnen adaptief gemaakt worden.

**Peer confrontation**:

We evalueren een systeem door het te laten concureren met een ander systeem. Er wordt een match gespeeld tussen meerdere systemen. Dit komt vaak voor in spellen en bij multi-agent research. Het resultaat, van soms meerdere metingen, wordt gebruikt om te bepalen welk systeem het beste is. Zoals duidelijk is, is het grootste probleem dat alle metingen relatief zijn. Ondanks dit relatieve karakter kan men nog steeds een gemiddelde performance bepalen. Het grootste probleem is robustheid en standaardisatie van de resultaten. Hoe bepaalt men de resultaten van twee "competities" als de deelnemers anders zijn? Gebruik rankings zoals Elo of meer geavanceerde systemen. Het beste is ook om competitie na competitie voorgaande tegenstanders terug mee te laten doen. Let wel op dat systemen niet gaan specialiseren om tegen bepaalde tegenstanders te spelen.

Een alternatief is om systemen tegen gestandaardiseerde tegenstanders te laten spelen. Hoe kiezen en hoe ook weer hetzelfde specialisatieprobleem tegen gaan? Voorgaande spelers met andere parameters? Tegenstanders laten cheaten met meer info?

**Highlights en toekomst**:

Een werkpunt is dat de evaluatie van AI scattered is over verschillende disciplines en er is veel duplicated effort. Hier een discipline op zich van maken, of een meer cross pollination tussen disciplines of centrale organisaties (NIST, DARPA, AAAI). NIST heeft hier vooral wat aan proberen te doen. Ook zijn er workshops georganiseerd voor AI evaluatie.

## 1.2 Naar ability-oriented evaluation

Task-oriented/narrow/specific AI is niet voldoende voor sommige toepassingen. Denk bijvoorbeeld aan cognitieve taken, artificiele huisdieren, assistenten, etc. Deze zijn niet gemaakt voor specifieke taken maar een variateit. Dus het systeem heeft een redenatievermogen nodig, inductief leervermogen, verbaal vermogen, bewegingsmogelijkheden, etc.

Als we terugkijken naar de definitie van Minsky moeten we kijken hoe een AI doet in een categorie: optimaal als het de beste is, super-strong human als het beter is dan alle mensen, super-human als het beter is dan de meeste mensen, par-human als het gelijk is met de meeste mensen en sub-human als het slechter is dan de meeste mensen. Dus AI is al veel vooruitgegaan. Super-human in de 19de eeuw voor berekeningen, in 1940 super-human in cryptografie, simpele spellen in de jaren 60, schaken in de jaren 90, speech recognition in de 2000, zeer veel gebieden in 2010s. Bij sommige toepassingen wordt ook de Turing scale gebruikt, waar par-human 0 is.

Realiseer ons dat nog geen systeem al deze dingen tegelijk kan leren, dus het is nog task-specific. Soms is een big-switch approach wel nuttig, zoals een systeem dat kan detecteren welk spel het nu aan het spelen is.

Nu gaan we bespreken hoe we een systeem per "ability" kunnen evalueren:

**Cognitive ability**:

Wat zijn cognitieve vaardigheden eigenlijk? Een eigenschap van indviduen dat hen toestaat om een tal van informatie-processing taken te doen. Deze vaardigheid is algemeen en slaat op een groot aantal taken. Bij general AI slaat deze vaardigheid op alle taken. Het grootste probleem is dat vaardigheden eigenschappen zijn en dus geconceptualiseerd en geidentificeerd moeten worden. Taken kunnen gezien worden als meetinstrumenten, maar vaardigheden zijn eerder constructies. In de psychologie zijn al een aantal cognitieve vaardigheden beschreven. In het handboek dat we gebruiken staan al een aantal gebieden: problem solving, use of knowledge, reasoning, learning, perception, natural language processing, etc..

Ability-oriented evaluation staat nog aan een beginnend stadium. Eerst omdat dit een zeer complex is. Ten tweede omdat de "taken" niet duidelijk omschreven zijn. Het hangt af van de conceptualisatie (weer het belang van theorien) en dan moet men representatieve oefeningen vinden. Ten derde zijn er nog niet heel veel general AI systemen. Dus zover leek task-oriented evaluation voldoende. Maar hier is nu verandering in aan het komen.

**Anthropocentrisme: psychometrics**:

Ontwikkeld eind 19de eeuw en de eerste helft 20ste eeuw. Verschil maken

tussen taken die specifieke vaardigheden nodig heeft tegenover algemene vaardigheden. Denk aan de "idiot savant" vs polymath (Narrow vs general AI). Kunnen we een IQ test aan een AI geven? Deze zijn al gestandaardiseerd, zijn abstract dus cultuur onafhankelijk. Er is heel wat kritiek op IQ tests. Men wilde Watson een IQ test laten doen met twee levels (om weer overfitting te voorkomen). Maar in 2003 had men al een klein programma geschreven (<1000 loc) dat goed scoorde op IQ tests, dus een IQ test is slecht en evalueert geen intelligentie. Dus zijn ze te antropocentrisch. Ondanks deze tekortkomingen zijn IQ tests toch populairder aan het worden voor AI evaluatie.

Hoe evalueert men nu dieren? Via rewards, zoals reinforcement learning. Dit gebruikt men dan voor comparative cognitive study. Misschien kan men reward systems gebruiken om AI te evalueren. Hoe taken en abilities selecteren? Dit moet systematisch gebeuren. Sommigen zijn te makkelijk (memory), sommigen te moeilijk (orientatie, herkenning en interactie in de wereld).

**Evaluatie met AIT**:

Als intelligentie als informatieverwerking kan beschouwd worden dan is het logisch dat men naar een formele basis en framework voor intelligentie kan zoeken. Dit is gedaan met algorithmic information theory (AIT). AIT wordt gekarakteriseerd door het definieren van tests vanuit formele information-based principes. Dit staat in contrast met andere technieken waar het vooral voortvloeit vanuit trial-and-error of op andere arbitraire manieren. Dit kan ook gebruikt worden in combinatie met andere technieken.

**Universal psychometrics**:

Deze techniek probeert universele tests te bekomen, universeel als in elke biologisch of artificieel systeem. Omdat niet alles in dezelfde taal of iets dergelijks communiceert moet dit adaptief zijn.

Een eerste voorbeeld is een omgeving die kan aanpassen aan de performance en snelheid van het systeem. Dit lijkt wat op difficulty-driven sampling, maar genereert de omgeving dus is adaptief. Deze omgeving kan gegenereerd worden vanuit AIT zoals boven vermeld.

**Highlights and directions of the evaluation of general-purpose AI systems**:

## 1.3 Conclusion

Summing up, AI requires an accurate, effective, non-anthropocentric, meaningful and computational way of evaluating its progress, by evaluating its

artefacts.

## 2 Notes

Wel: "Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement" -> *We identify three kinds of evaluation: Human discrimination, problem benchmarks and peer confrontation.* [2] `https://sci-hub.mksa.top/10.1007/s10462-016-9505-7`

"various video game platforms for AI evaluation" -> `https://ojs.aaai.org//index.php/aimagazine/article/view/2748`

Measuring progress: `https://www.eff.org/ai/metrics`

[1]

[4]

[3]

## References


[1] N. Alvarado, S.S. Adams, S. Burbeck, and C. Latta. "Beyond the Turing test: Performance metrics for evaluating a computer simulation of the human mind". In: *Proceedings 2nd International Conference on Development and Learning. ICDL 2002* (). DOI: 10.1109/devlrn.2002.1011826 (cit. on p. 8).

[2] José Hernández-Orallo. "Evaluation in artificial intelligence: From task-oriented to ability-oriented measurement". In: *Artificial Intelligence Review* 48.3 (2016), pp. 397–447. DOI: 10.1007/s10462-016-9505-7 (cit. on p. 8).

[3] Philip Hingston. "A new design for a Turing test for Bots". In: *Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games* (2010). DOI: 10.1109/itw.2010.5593336 (cit. on p. 8).

[4] Paul Schweizer. "The Truly Total Turing Test". In: *Minds and Machines* 8.2 (1998), pp. 263–272. DOI: 10.1023/a:1008229619541 (cit. on p. 8).