

From Neurons to Public-Private Labs

How Testing and Evaluation Determine AI Adoption

Maarten de Rijke

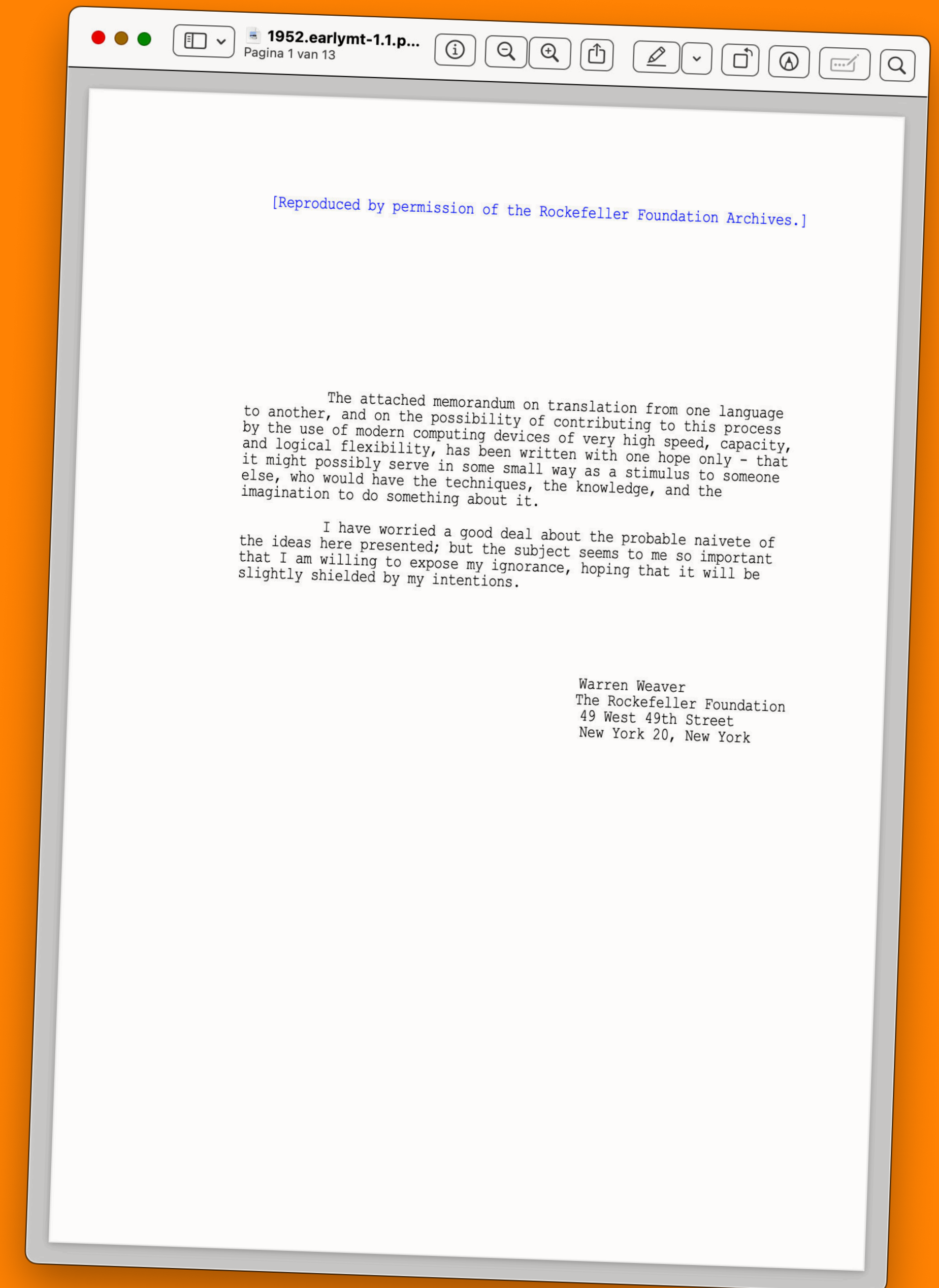
University of Amsterdam & ICAI

m.derijke@uva.nl

A look back

“Solve” natural language processing
(or at least **machine translation**)
as a summer project

1949



Another look back

Artificial intelligence as a summer project

“Solve” AI in a summer

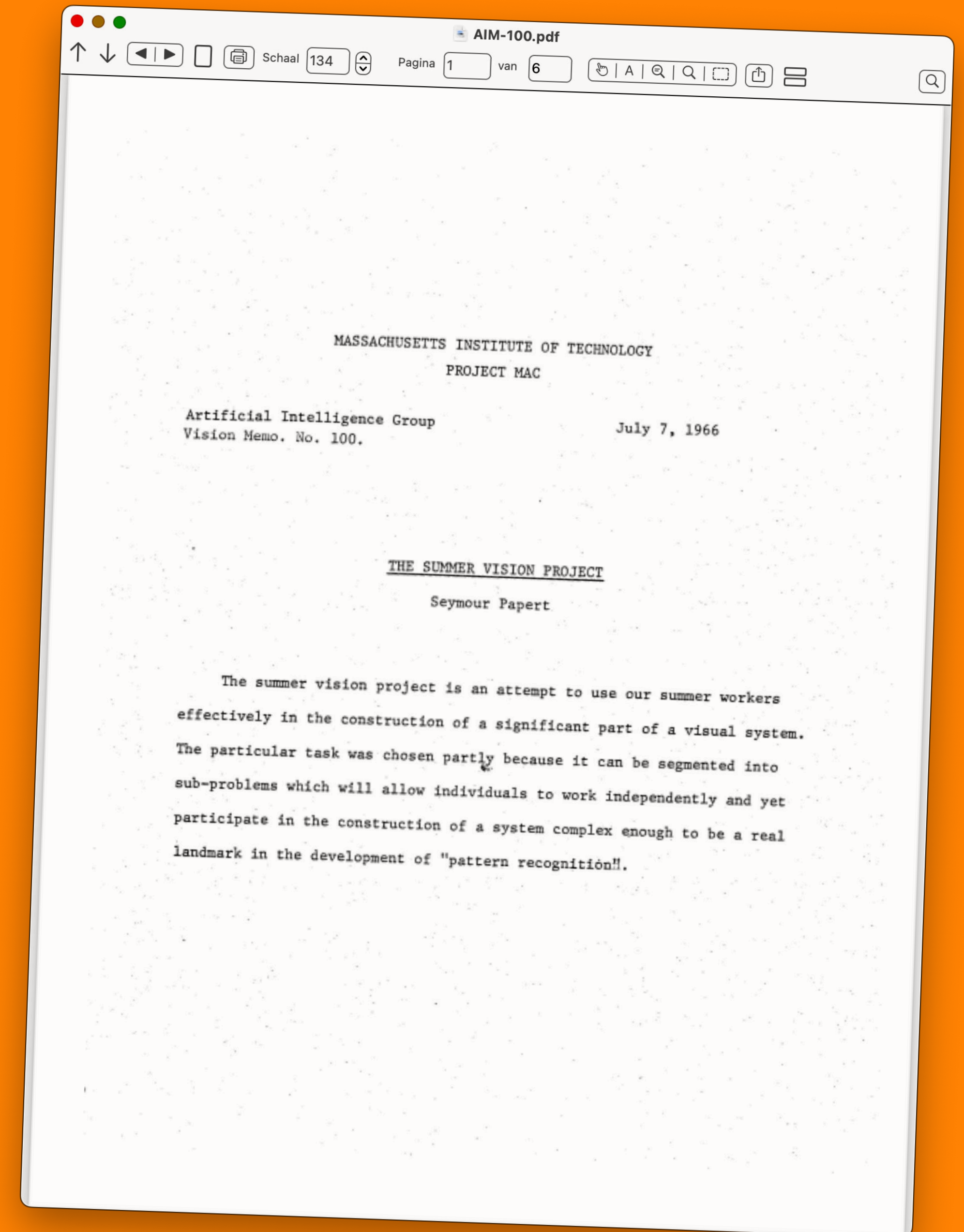
1955



And one more

“Solve” computer vision
as a summer project

1966



Fast forward to today

AI is overhyped vs. AI is accelerating uncontrollably vs....

**From individual neurons to multi-agent systems:
a recurring pattern of abstraction**

neurons/weights → **layers**
→ **architectures**
→ **models-as-functions**
→ **training objectives**
→ **natural language as interface**
→ **tool use**
→ **agents**
→ **multi-agent systems**

Setting the scene

At what point does an AI capability become something an organization or a person is willing to rely on?

And what has to happen between now and then?

**Starting from search and recommendation:
abstraction in a digital setting**

Classical information retrieval and recommender systems

- Single-shot query → ranked list
- Static, session-independent preference model
- Evaluation ends at the ranking

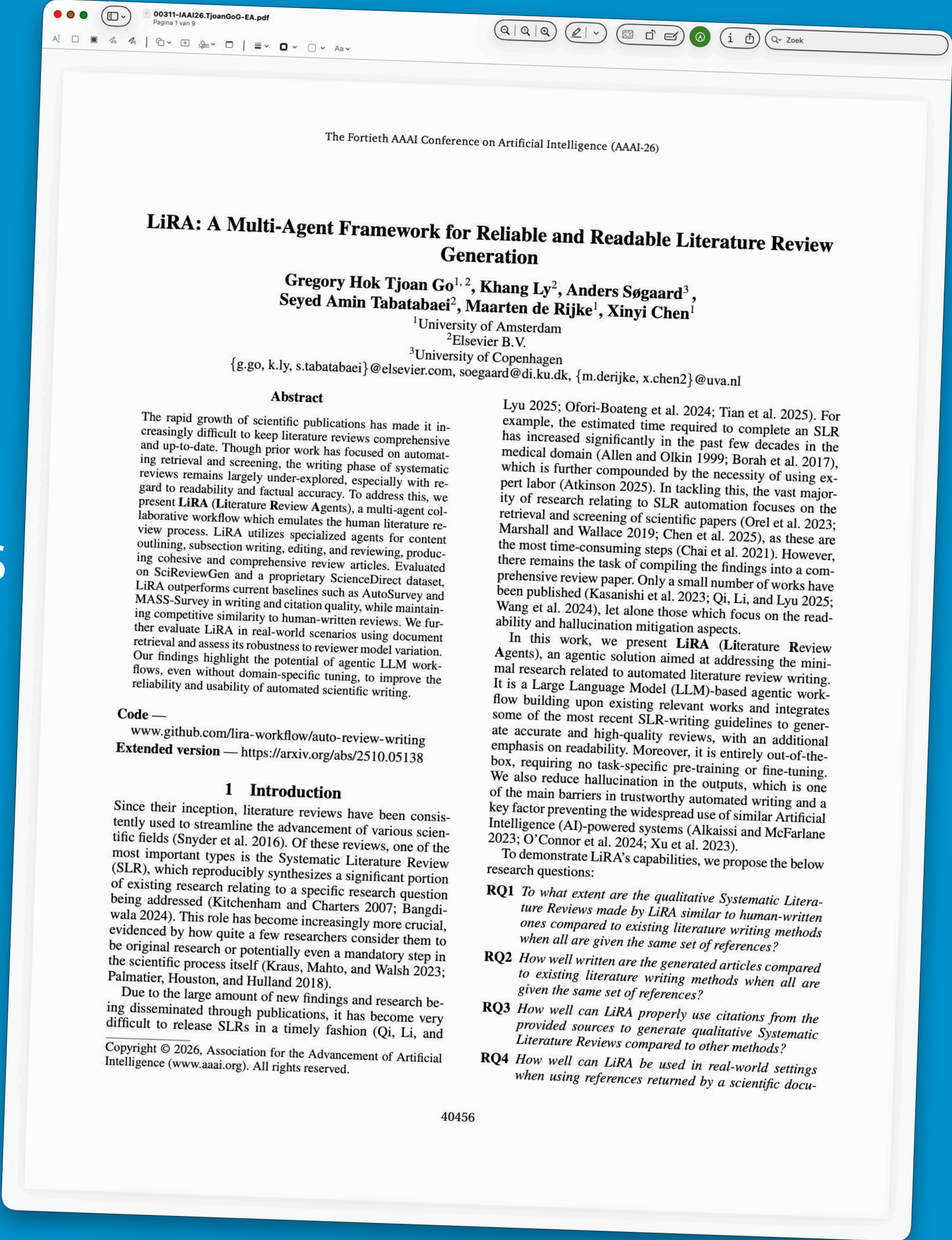
Agentic information retrieval and recommender systems

- Multi-step planning and synthesis
- Interactive, in-session negotiation
- Can act, not just suggest (book, buy, subscribe)

Example

Automatically generate systematic reviews

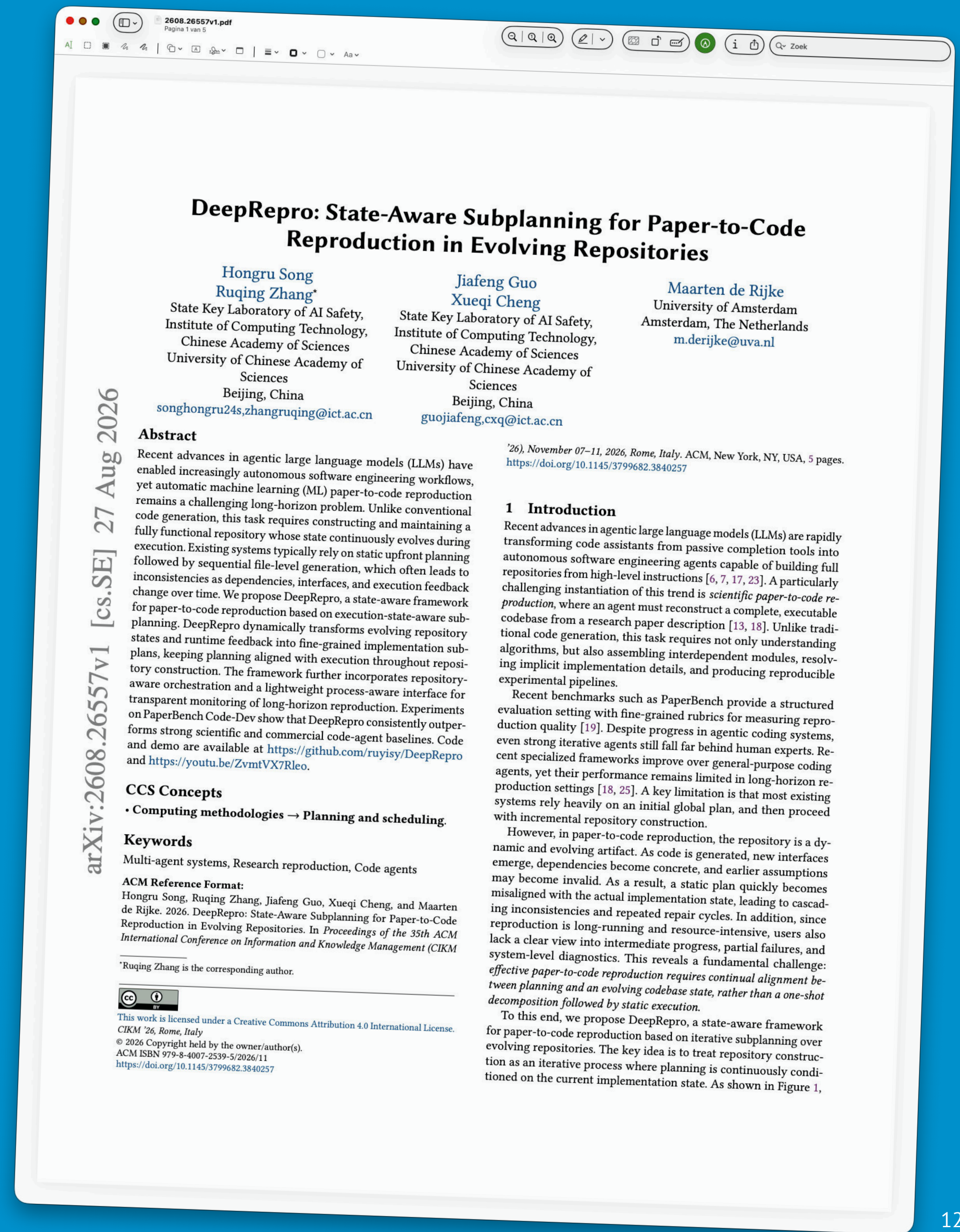
January 2026



One more example

Have agents reproduce a paper, using online code repositories

November 2026



Some of today's research challenges in search & recommendation agents

1. Planning & long-horizon reasoning

Compounding errors across steps

2. Memory & state management

What to keep, for how long

3. Tool-use reliability

A malformed call propagating downstream

4. Evaluation

Trajectory-dependent success resists static benchmarks

5. Safety & interruptibility

The digital red button

6. Multi-agent coordination

Collusion, deadlock, propagated hallucination

7. Cost & latency

Agent loops multiply model calls

1. Query decomposition

As an explicit exploration–exploitation trade-off

2. Clarifying questions

As a scoreable action

3. Answer-reconstruction search density

Measuring compressed search effort

4. Can we trust the referee?

Re-rankers as relevance judges

- 1. Joint knowledge selection + generation**
Direct retrieval-augmented optimization
- 2. Calibrated uncertainty**
Probabilistic reasoning
- 3. Continual pre-training against staleness**
CorpusBrain++
- 4. Correcting how evidence is weighted**
Self-adaptive cognitive debiasing

**Abstraction, extended to embodiment,
where the stakes change**

An example

www.ft.com/content/5c2ae092-14af-4bd1-8e3d-8996755cba6a?syn-25a6b1a6=1

FT CuspAI's Max Welling: 'We are building molecules to remove forever chemicals from water'

FINANCIAL TIMES

Subscribe Sign In

HOME WORLD US COMPANIES TECH MARKETS CLIMATE OPINION LEX WORK & CAREERS LIFE & ARTS HOW TO SPEND IT

FT Collections AI Exchange

Show articles

Artificial intelligence + Add to myFT

CuspAI's Max Welling: 'We are building molecules to remove forever chemicals from water'

The co-founder of the UK-based science start-up explains how AI can help create new materials to address some of the world's most complex challenges



X f in Save

John Thornhill

Published JUL 30 2026

14

Stay informed with free updates

Simply sign up to the Artificial intelligence myFT Digest -- delivered directly to your inbox.

Enter your email address

Sign up

Max Welling is a computer science professor at the University of Amsterdam and the co-founder of CuspAI, a UK-based start-up aiming to revolutionise the discovery of new materials. The company's pitch is that it will use AI to achieve breakthroughs in "months, not millennia".

He has flitted between the worlds of research and commerce for three decades. His academic life includes posts at University College London, University of Toronto, University of California, Irvine and California Institute of Technology, while his business career ranges from a stint as a start-up founder as well as roles at Qualcomm and Microsoft.

Welling co-founded CuspAI with Chad Edwards, a former executive at Quantinuum, in 2024, to create a "search engine" for materials science. The company has assembled a heavyweight board of advisers, including the AI researchers [Geoff Hinton](#) and [Yann LeCun](#), Lord John Browne, the former chief executive of BP, and Martin van den Brink, the former president of ASML.

1. Multimodal backbones

Output robot, machine and physical actions instead of text

2. Language interface

Becomes the human-robot interface

3. Tool use

Becomes embodied tool use

4. Agent loop

Becomes an embodied control loop

5. Multi-agent

Becomes robot fleets / swarms

1. Irreversibility replaces resampling

A dropped payload isn't a regenerated token

2. Perception becomes noisy and adversarial

Sensor readings cover only part of the scene and their calibration degrades over time (drift), so the same object can appear different to the sensor today than it did yesterday

3. Actions can cause harm independent of intent

A wrong torque command can injure a person or damage equipment on execution

The pattern across domains

Domain	Generative layer	Verification mechanism
Manufacturing	Learned manipulation policies	Force limiting (ISO/TS 15066)
Engineering	Generative design	Finite element analysis
Materials science	Candidate generation	Self-driving labs
Drug design	Generative molecule models	Multi-phase clinical trials
Healthcare	Diagnostic classifiers	Regulatory clearance + clinician oversight
Climate science	Learned weather emulators	Calibration in the tails
Water infrastructure	RL-based network control	Fail-safe defaults

The two sentence message

In every domain, outputs from generative models are checked by a separate process before they take effect: simulation, clinical trials, physical interlocks, or regulatory review

Which check is used depends on how well the underlying system is understood and how accurately it can be simulated

Does better evaluation change anything at scale?

Does this matter at scale?

Three different speeds

- Methods: lab capability
- Applications: what's built on top
- Adoption: what's actually trusted and deployed

Autonomy: to which degree can you do all three on your own terms?

Key points

- Adoption is almost always the slowest
- Adoption is almost always the real bottleneck
- Treat AI as a normal technology

Depending on others for methods or applications limits how independently adoption decisions can be made

The adoption bottleneck put differently

Adoption is the grounding problem, but now scaled to a society

A capability in a lab is worthless until an institution builds the trust to use it

1. State-of-the-art agents chase accuracy, ignore cost

Needlessly complex, expensive agents

2. Model-level vs. application-level needs conflated

Different audiences, same benchmark

3. A single success metric hides operational flaws

Consistency, sensitivity, degradation

Research + applications + evaluation performed in context determine whether an AI solution gets adopted

Develop AI knowledge, push AI innovation, build AI talent and decide on adoption on (y)our own terms

This is the challenge that ICAI, the Innovation Center for AI, helps address

Distributed and collaborative AI power in the Netherlands

The challenge

AI will reshape every sector of society. The critical question is who controls it. ICAI was founded in 2018 on the belief that in an open society, technological power must be distributed, not concentrated in the hands of a few.

The model

Multi-stakeholder AI labs connecting universities, companies, government, and society. Each lab runs five years, with at least five PhD students, embedded on-site at partner organizations, creating a continuous flow of knowledge and talent between academia and practice.

Key facts

Eight domains

Strong institutions

Mobility & logistics

Services

Industry & infrastructure

Agriculture & food

Knowledge & education

Sustainability

60

Labs

670

AI talents

161

Partners

€307M Invested, just in talent

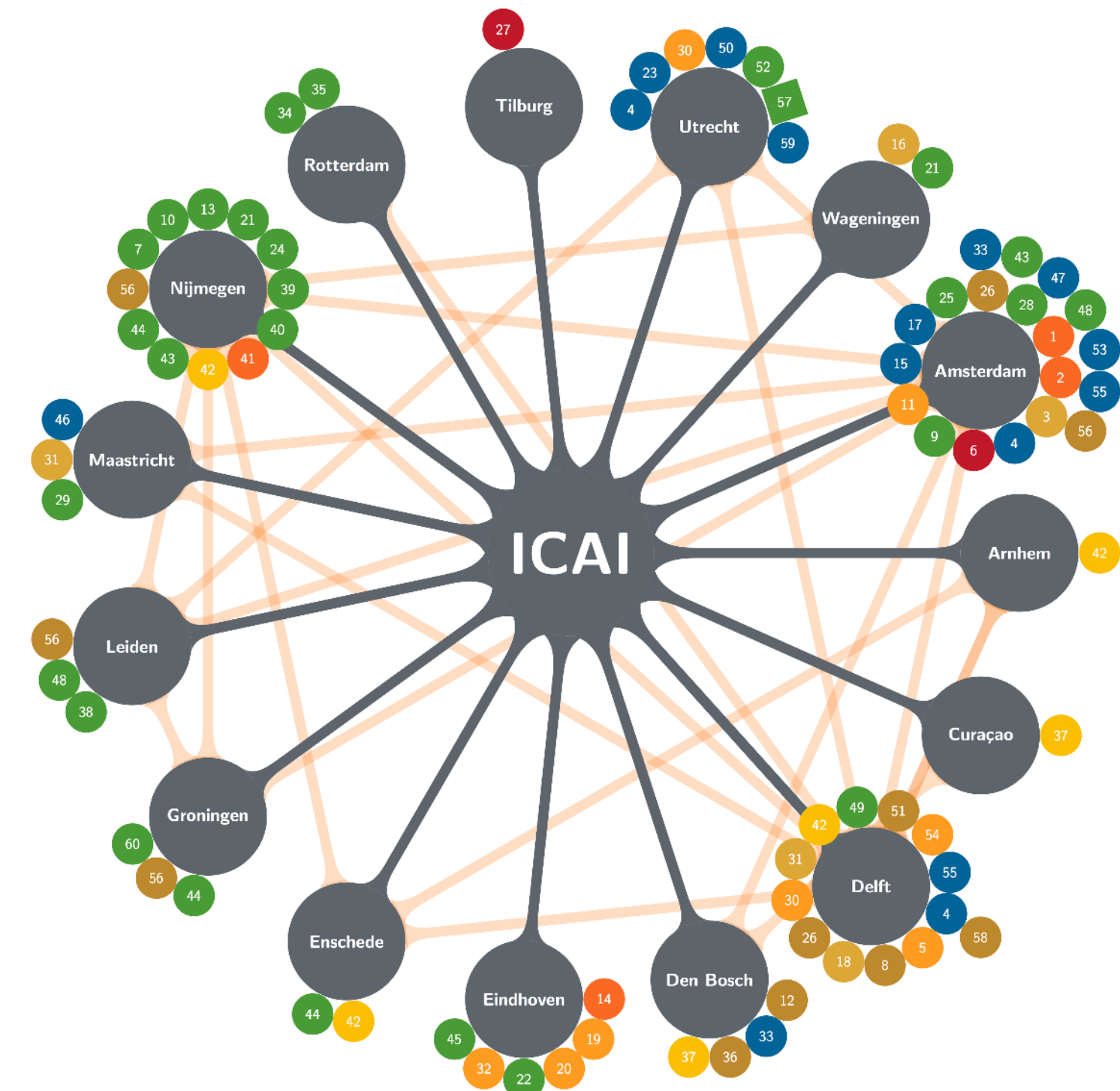
5000 SMEs advised

6000 Publications

300 Co-productions

85% Talent retention

15 Locations, nationwide



Building “soft infrastructure” for use-inspired research

Beyond the individual labs

Sharing lessons & Best practices

ICAI systematically captures and disseminates what works, across labs, universities, and sectors. Through events, publications, direct exchange, we help the broader Dutch AI community learn faster and avoid duplication of effort.

Large-scale collaborative grants

ICAI helps set up or run large-scale programs, such as ROBUST, a €87.5M NWO co-funded long-term program, running 2022–2027 with 17 labs. Phase 2 (2027–2032) is in preparation with 13 newly selected labs.

Shaping the national agenda

ICAI helps shape national AI policy through consultation with Eerste Kamer, Tweede Kamer, Economic Affairs and Climate. ICAI also engages with underrepresented regions to distribute AI opportunities as widely as possible.

Building “soft infrastructure” for use-inspired research

International advisory roles

ICAI’s Dutch AI model is in demand internationally. We advise organizations worldwide on building multi-stakeholder AI ecosystems modelled on the Dutch approach. 🇦🇺 · 🇫🇮 · 🇩🇪 · 🇬🇷 · 🇮🇳 · 🇭🇺 · 🇳🇴 · 🇸🇪

Frontier directions

Defense & sovereignty

Responsible AI for a resilient democracy

Health

Precision medicine, care pathways, reduce admin

Semi-conductor industry

AI for manufacturing & engineering

Target by late 2027

80

Labs

800

AI talents

175

Partners

Our unique position

ICAI occupies TRL 1–7, between blue-sky research and deployment, a middle ground that no one else fills. We complement high-risk, long-term investments and short-term market implementation initiatives.

The ROBUST program

10 year long-term program run out of ICAI
Trustworthy AI-based systems for sustainable growth

ROBUST in numbers



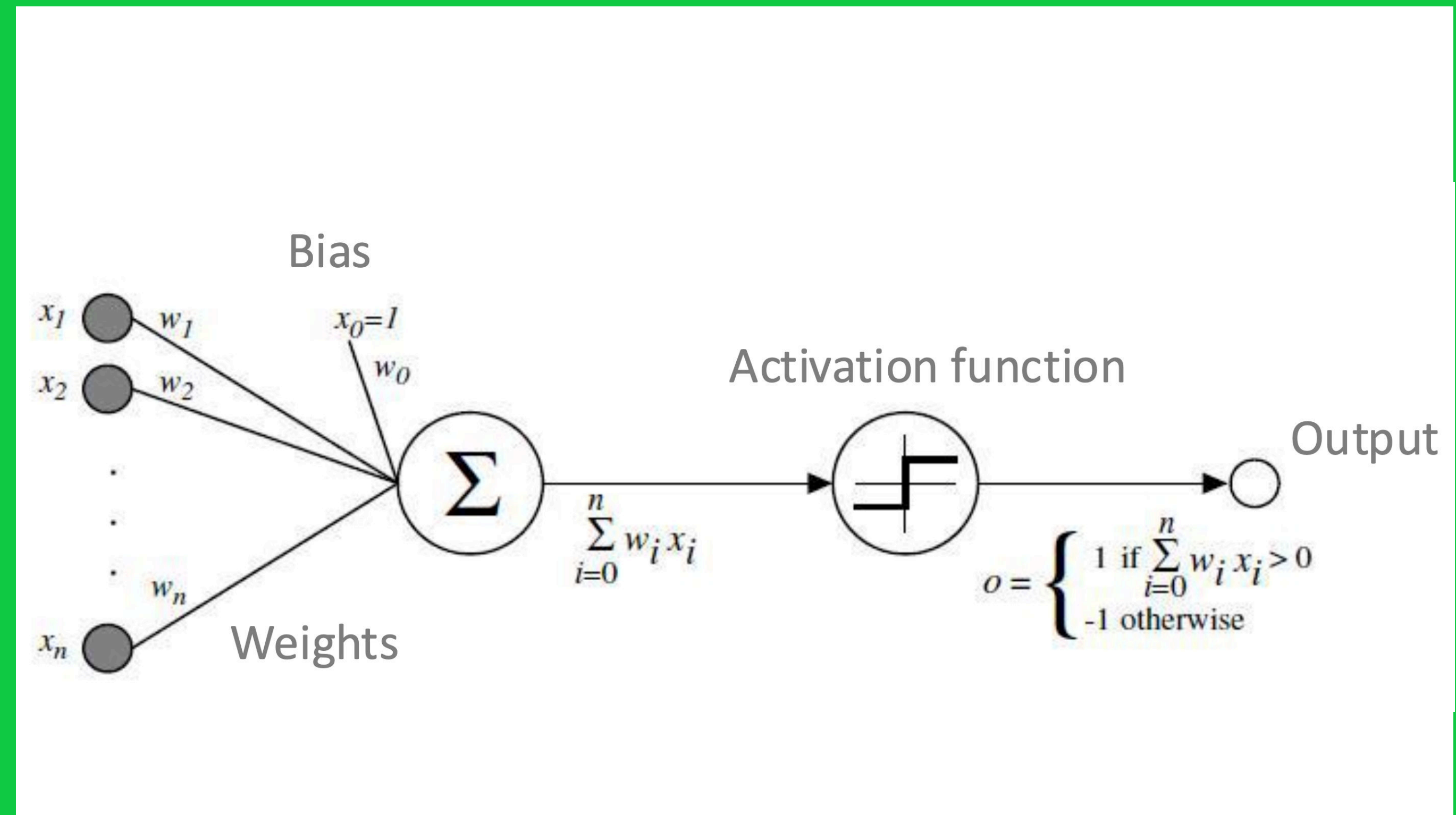
Let's wrap up

One more historical example

Perceptron: A machine for binary classification

Core idea

- One weight per input
- Multiply weights with input and add bias term
- If result exceeds threshold, return 1, else -1



It was noticed, fast

NEW NAVY DEVICE LEARNS BY DOING

Psychologist Shows Embryo
of Computer Designed to
Read and Grow Wiser

WASHINGTON, July 7 (UPI)
—The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The Navy's \$2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's demonstration for newsmen.

The service said it would use this principle to build the first of its Perceptron thinking machines that will be able to read and write. It is expected to be finished in about a year at a cost of \$100,000.

Dr. Frank Rosenblatt, designer of the Perceptron, conducted the demonstration. He said the machine would be the first device to think as the human brain. As do human beings, Perceptron will make mistakes at first, but will grow wiser as it gains experience, he said.

Dr. Rosenblatt, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said Perceptrons might be fired to the planets as mechanical space explorers.

Without Human Controls

The Navy said the perceptron would be the first non-living mechanism "capable of receiving, recognizing and identifying its surroundings without any human training or control."

The "brain" is designed to remember images and information it has perceived itself. Ordinary computers remember only what is fed into them on punch cards or magnetic tape.

Later Perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech or writing in another language, it was predicted.

Mr. Rosenblatt said in principle it would be possible to build brains that could reproduce themselves on an assembly line and which would be conscious of their existence.

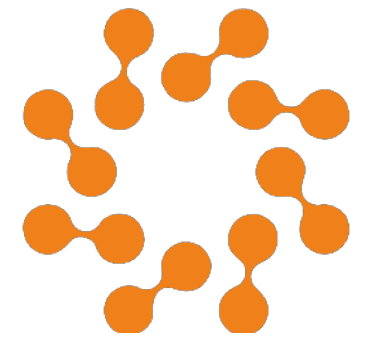
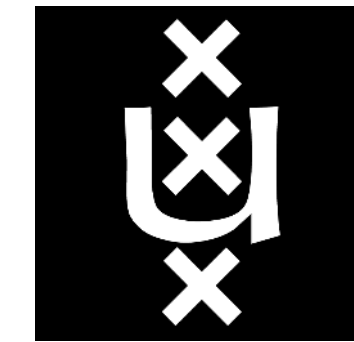
In today's demonstration, the "704" was fed two cards, one with squares marked on the left side and the other with squares on the right side.

Learns by Doing

In the first fifty trials, the machine made no distinction between them. It then started registering a "Q" for the left squares and "O" for the right squares.

Dr. Rosenblatt said he could explain why the machine learned only in highly technical terms. But he said the computer had undergone a "self-induced change in the wiring diagram."

The first Perceptron will have about 1,000 electronic "association cells" receiving electrical impulses from an eye-like scanning device with 400 photo-cells. The human brain has 10,000,000,000 responsive cells, including 100,000,000 connections with the eyes.



From Neurons to Public-Private Labs

How Testing and Evaluation Determine AI Adoption

Maarten de Rijke

University of Amsterdam & ICAI

m.derijke@uva.nl