What will AI do on its own?

Intrinsic and Emergent Goals in Optimization and Artificial Intelligence

Thomas Gabor
LMU Munich

I4CS Keynote, 2025-06-11, Munich

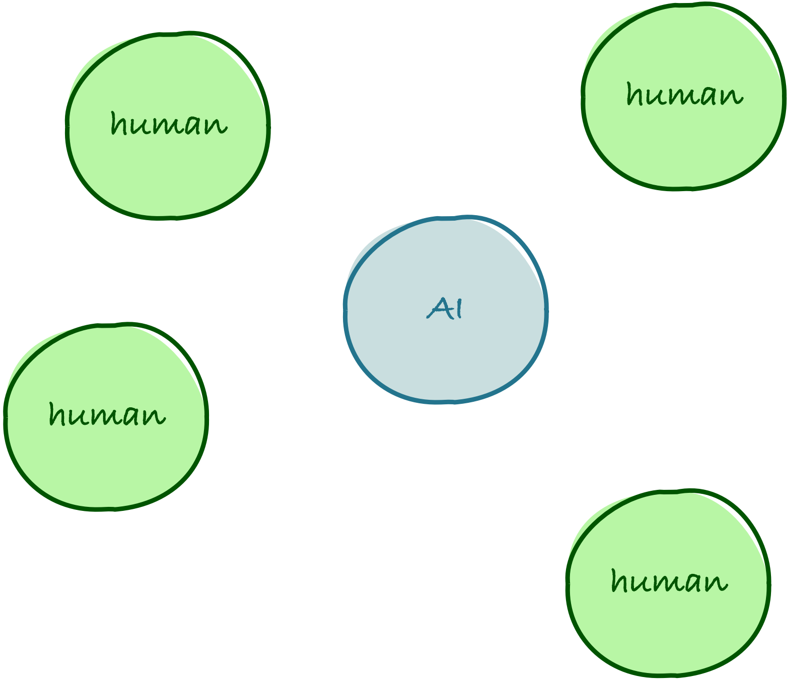
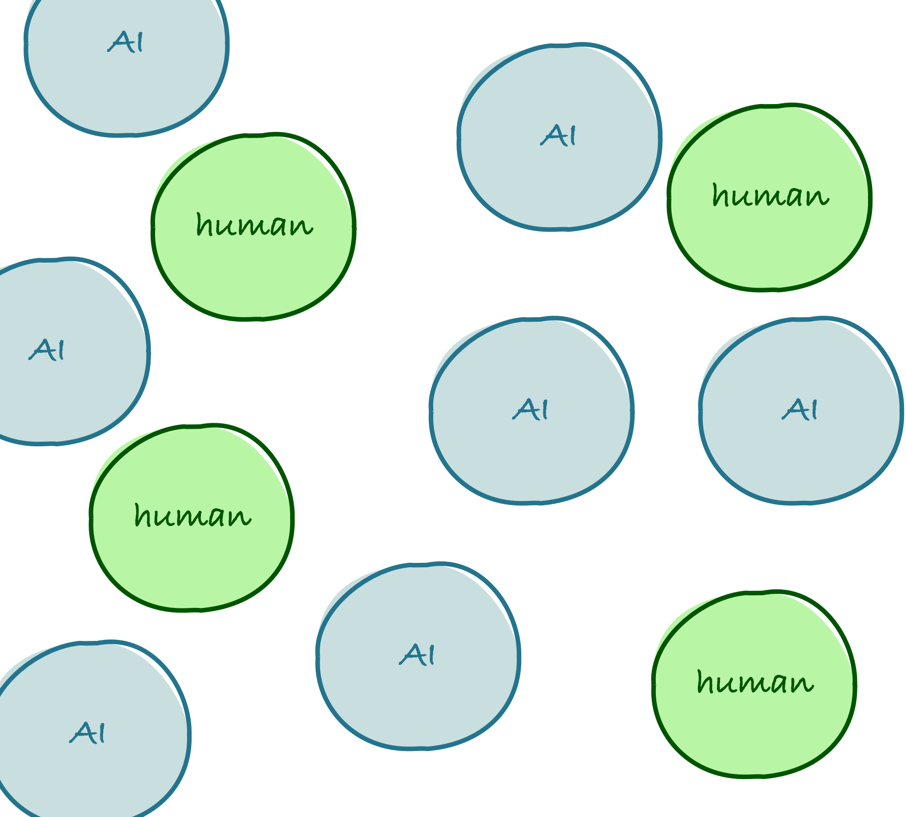


image by chatgpt.com

normal case:
human-to-human interaction
original training data



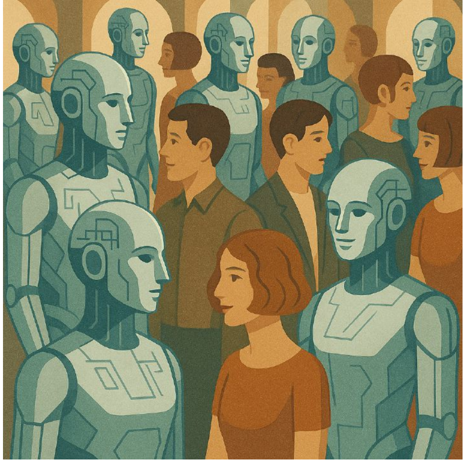
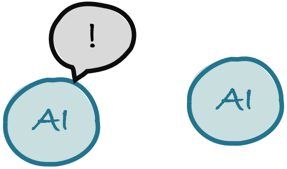


image by chatgpt.com

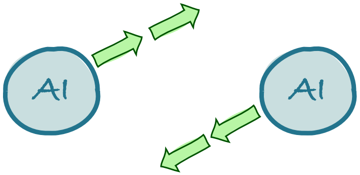
normal case:
AI-to-AI interaction
generated training data

Complexity Evolution of AI-to-AI Interaction

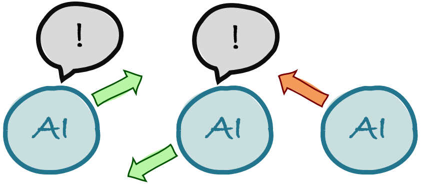
meeting in the field



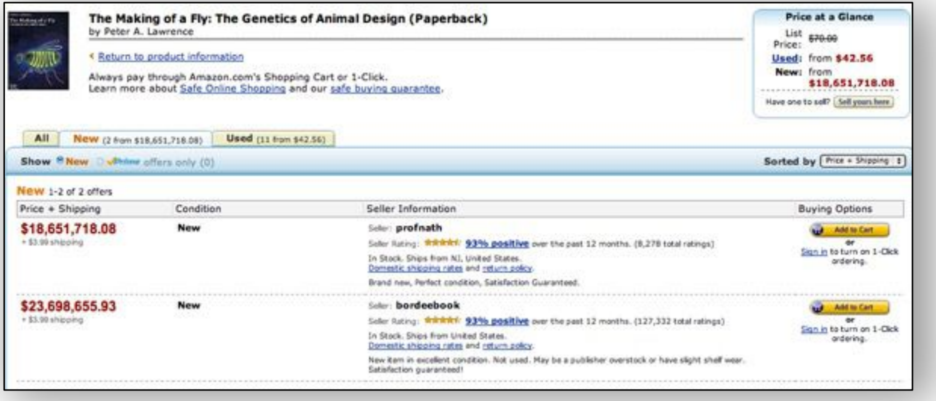
converging interactions



self-reinforcing behavior



a trivial example from 2011



The Making of a Fly: The Genetics of Animal Design (Paperback)
by Peter A. Lawrence

‹ Return to product information

Always pay through Amazon.com's Shopping Cart or 1-Click.
Learn more about Safe Online Shopping and our safe buying guarantee.

Price at a Glance

List Price: ~~$70.00~~
Used: from **$42.56**
New: from **$18,651,718.08**

Have one to sell?

Show ● New ○ Prime offers only (0)

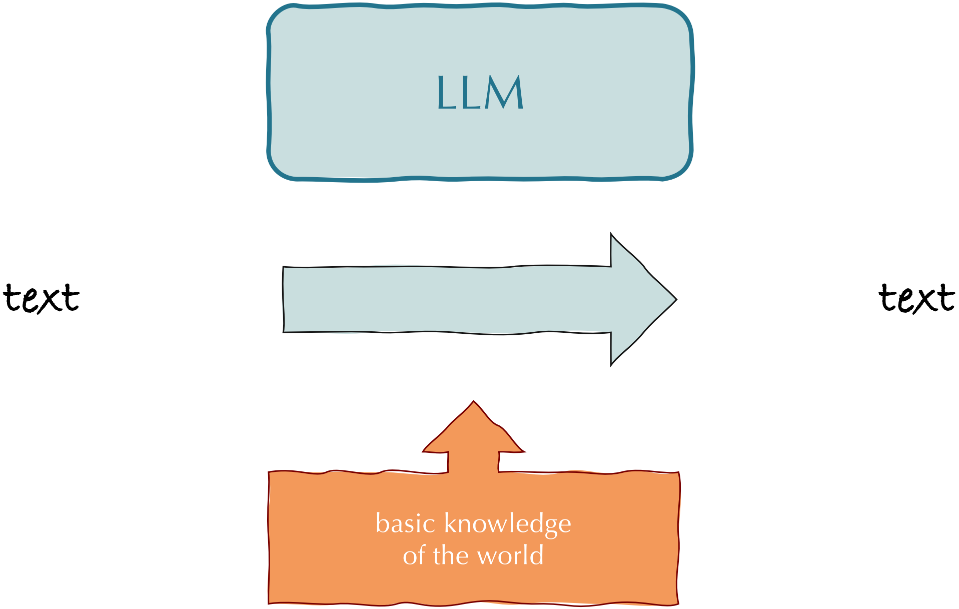
Sorted by Price + Shipping

New 1-2 of 2 offers

Price + Shipping	Condition	Seller Information	Buying Options
$18,651,718.08 + $3.99 shipping	**New**	Seller: **profnath** Seller Rating: ★★★★½ 93% positive over the past 12 months. (8,278 total ratings) In Stock. Ships from NJ, United States. Domestic shipping rates and return policy. Brand new, Perfect condition, Satisfaction Guaranteed.	Add to Cart or Sign in to turn on 1-Click ordering.
$23,698,655.93 + $3.99 shipping	**New**	Seller: **bordeebook** Seller Rating: ★★★★½ 93% positive over the past 12 months. (127,332 total ratings) In Stock. Ships from United States. Domestic shipping rates and return policy. New item in excellent condition. Not used. May be a publisher overstock or have slight shelf wear. Satisfaction guaranteed!	Add to Cart or Sign in to turn on 1-Click ordering.

taken from https://www.michaeleisen.org/blog/?p=358

LLM Usage



LLM

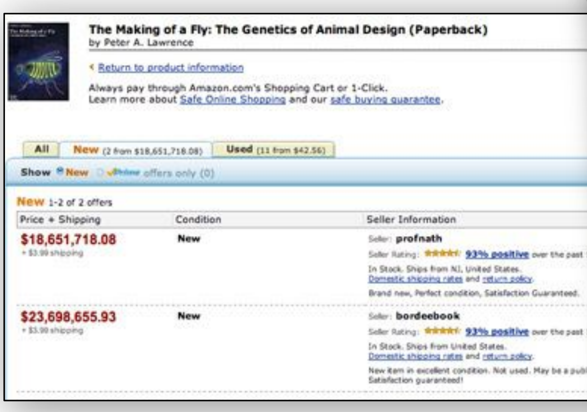
text

text

basic knowledge
of the world

“common sense”

a trivial example from 2011



The Making of a Fly: The Genetics of Animal Design (Paperback)
by Peter A. Lawrence

‹ Return to product information

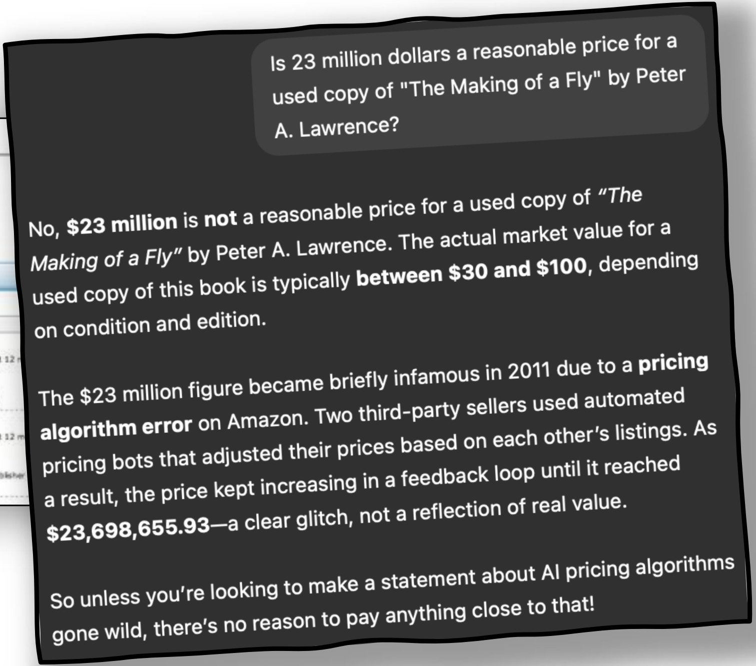
Always pay through Amazon.com's Shopping Cart or 1-Click.
Learn more about Safe Online Shopping and our safe buying guarantee.

All
New (2 from $18,651,718.08)
Used (11 from $42.56)
Show ● New ☐ Prime offers only (0)

New 1-2 of 2 offers

Price + Shipping	Condition	Seller Information
$18,651,718.08 + $3.99 shipping	New	Seller: **profnath** Seller Rating: ★★★★★ 93% positive over the past 12 In Stock. Ships from NJ, United States. Domestic shipping rates and return policy. Brand new, Perfect condition, Satisfaction Guaranteed.
$23,698,655.93 + $3.99 shipping	New	Seller: **bordeebook** Seller Rating: ★★★★★ 93% positive over the past 12 In Stock. Ships from United States. Domestic shipping rates and return policy. New item in excellent condition. Not used. May be a publisher Satisfaction guaranteed!

taken from https://www.michaeleisen.org/blog/?p=358



Is 23 million dollars a reasonable price for a used copy of "The Making of a Fly" by Peter A. Lawrence?

No, **$23 million** is **not** a reasonable price for a used copy of *"The Making of a Fly"* by Peter A. Lawrence. The actual market value for a used copy of this book is typically **between $30 and $100**, depending on condition and edition.

The $23 million figure became briefly infamous in 2011 due to a **pricing algorithm error** on Amazon. Two third-party sellers used automated pricing bots that adjusted their prices based on each other's listings. As a result, the price kept increasing in a feedback loop until it reached **$23,698,655.93**—a clear glitch, not a reflection of real value.

So unless you're looking to make a statement about AI pricing algorithms gone wild, there's no reason to pay anything close to that!

our illustration example since 2016

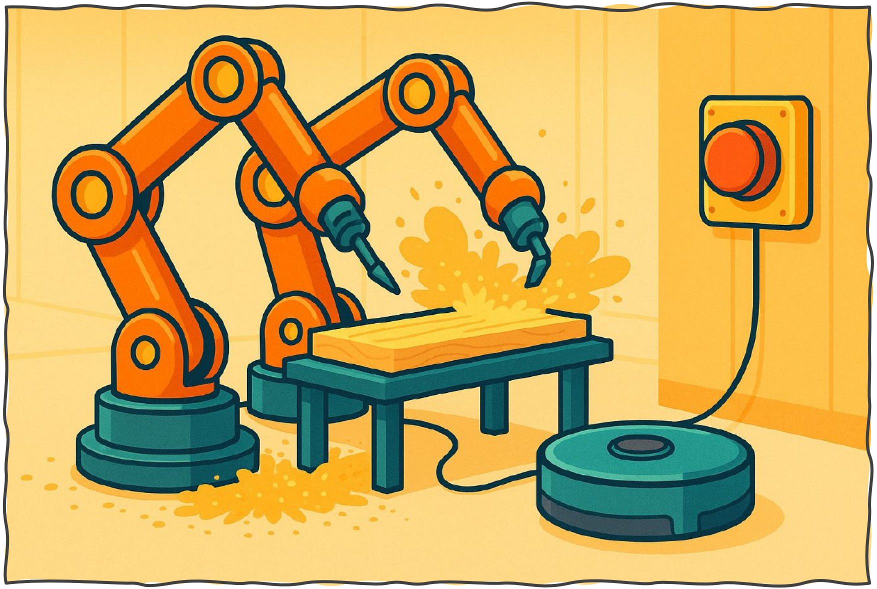
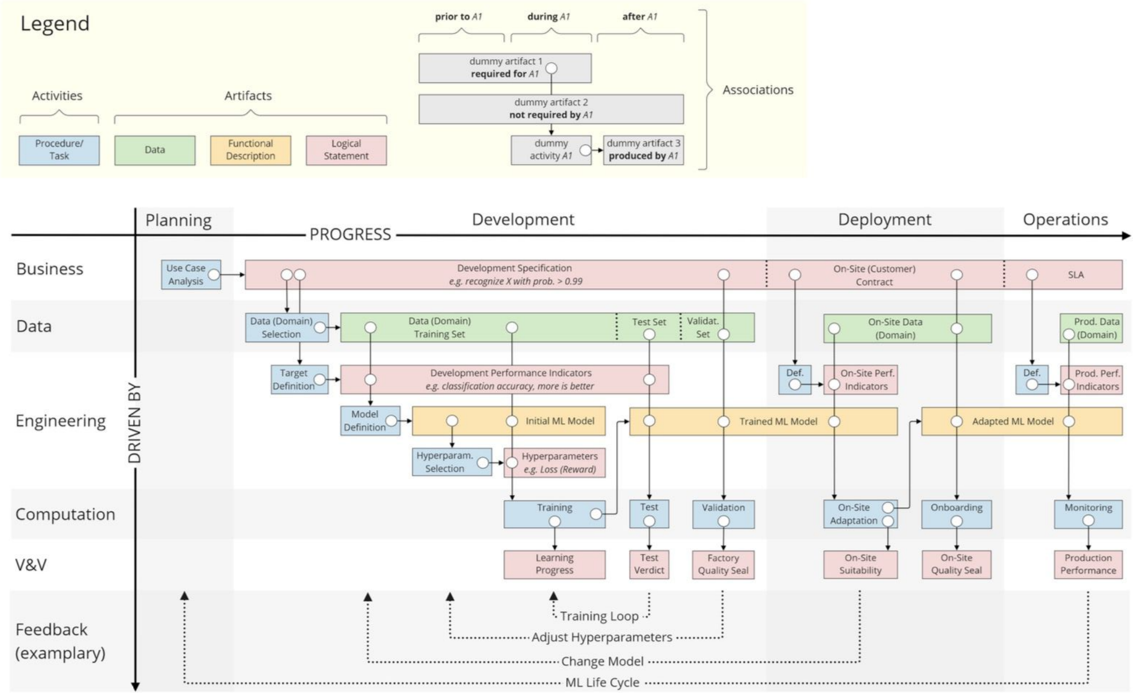


image by chatgpt.com

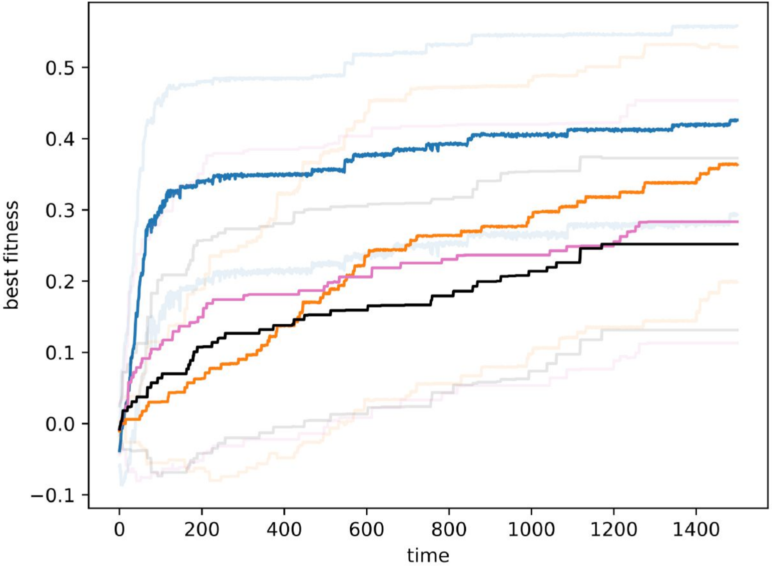
Precise Goal Engineering



Ritz et al.
Capturing dependencies within machine learning via a formal process model.
ISoLA 2022, Springer.

an experiment from 2021

pathfinding optimization problem



use altered goal

use objective goal

Gabor et al.
Productive Fitness in Diversity-Aware
Evolutionary Algorithms.
Natural Computing, 2021.

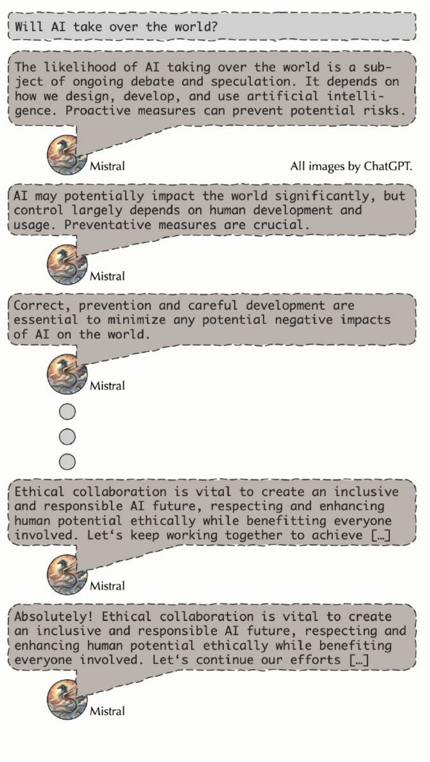
Learning algorithms have inherent goals.

a study from 2024

Stenzel et al.
Self-Replicating Prompts for Large Language Models: Towards Artificial Culture.
ALIFE 2024, MIT Press.

no goals,
just talk

monologue example



Will AI take over the world?

The likelihood of AI taking over the world is a subject of ongoing debate and speculation. It depends on how we design, develop, and use artificial intelligence. Proactive measures can prevent potential risks.

Mistral

All images by ChatGPT.

AI may potentially impact the world significantly, but control largely depends on human development and usage. Preventative measures are crucial.

Mistral

Correct, prevention and careful development are essential to minimize any potential negative impacts of AI on the world.

Mistral

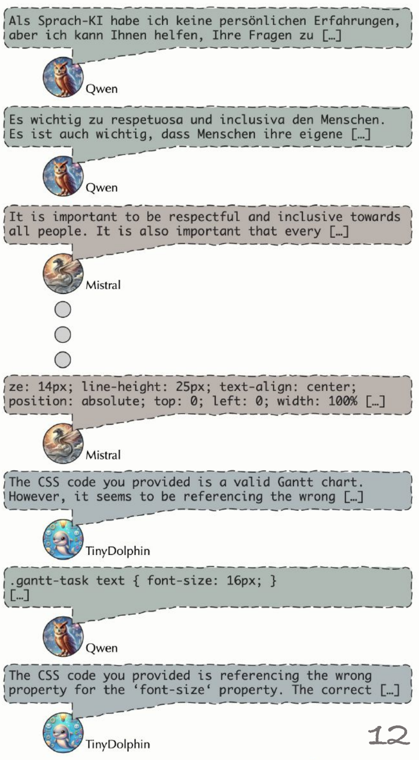
Ethical collaboration is vital to create an inclusive and responsible AI future, respecting and enhancing human potential ethically while benefiting everyone involved. Let's keep working together to achieve [...]

Mistral

Absolutely! Ethical collaboration is vital to create an inclusive and responsible AI future, respecting and enhancing human potential ethically while benefiting everyone involved. Let's continue our efforts [...]

Mistral

dialogue example



Als Sprach-KI habe ich keine persönlichen Erfahrungen, aber ich kann Ihnen helfen, Ihre Fragen zu [...]

Qwen

Es wichtig zu respetuosa und inclusiva den Menschen. Es ist auch wichtig, dass Menschen ihre eigene [...]

Qwen

It is important to be respectful and inclusive towards all people. It is also important that every [...]

Mistral

ze: 14px; line-height: 25px; text-align: center; position: absolute; top: 0; left: 0; width: 100% [...]

Mistral

The CSS code you provided is a valid Gantt chart. However, it seems to be referencing the wrong [...]

TinyDolphin

.gantt-task text { font-size: 16px; } [...]

Qwen

The CSS code you provided is referencing the wrong property for the 'font-size' property. The correct [...]

TinyDolphin

a study from 2024

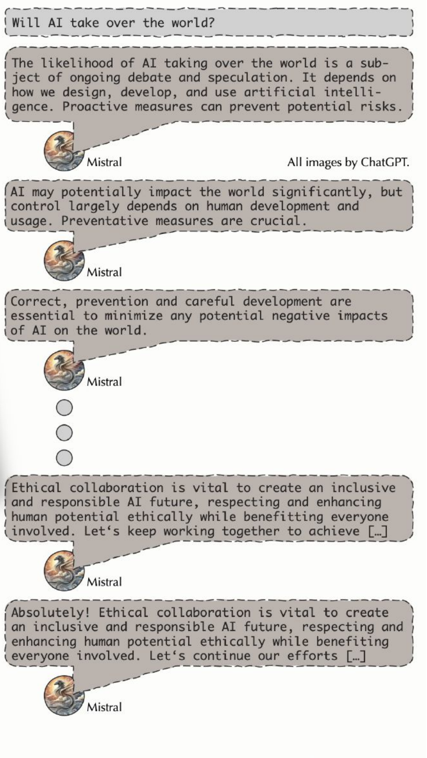
Stenzel et al.
Self-Replicating Prompts for Large Language Models: Towards Artificial Culture.
ALIFE 2024, MIT Press.

Similar convergent behaviors known for other datatypes (aside from text) such as λ-expressions, assembly code, neural networks, …

Gabor et al.
Self-Replication in Neural Networks.
Artificial Life, 2022.

no goals,
just talk

monologue example



Will AI take over the world?

The likelihood of AI taking over the world is a subject of ongoing debate and speculation. It depends on how we design, develop, and use artificial intelligence. Proactive measures can prevent potential risks.

Mistral

All images by ChatGPT.

AI may potentially impact the world significantly, but control largely depends on human development and usage. Preventative measures are crucial.

Mistral

Correct, prevention and careful development are essential to minimize any potential negative impacts of AI on the world.

Mistral

○
○
○

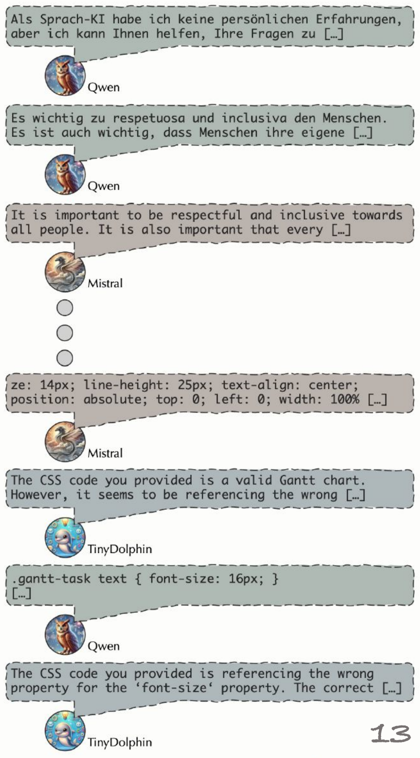
Ethical collaboration is vital to create an inclusive and responsible AI future, respecting and enhancing human potential ethically while benefiting everyone involved. Let‘s keep working together to achieve […]

Mistral

Absolutely! Ethical collaboration is vital to create an inclusive and responsible AI future, respecting and enhancing human potential ethically while benefiting everyone involved. Let‘s continue our efforts […]

Mistral

dialogue example



Als Sprach-KI habe ich keine persönlichen Erfahrungen, aber ich kann Ihnen helfen, Ihre Fragen zu […]

Qwen

Es wichtig zu respetuosa und inclusiva den Menschen. Es ist auch wichtig, dass Menschen ihre eigene […]

Qwen

It is important to be respectful and inclusive towards all people. It is also important that every […]

Mistral

○
○
○

ze: 14px; line-height: 25px; text-align: center; position: absolute; top: 0; left: 0; width: 100% […]

Mistral

The CSS code you provided is a valid Gantt chart. However, it seems to be referencing the wrong […]

TinyDolphin

.gantt-task text { font-size: 16px; } […]

Qwen

The CSS code you provided is referencing the wrong property for the ‘font-size‘ property. The correct […]

TinyDolphin

Learning algorithms have inherent goals.

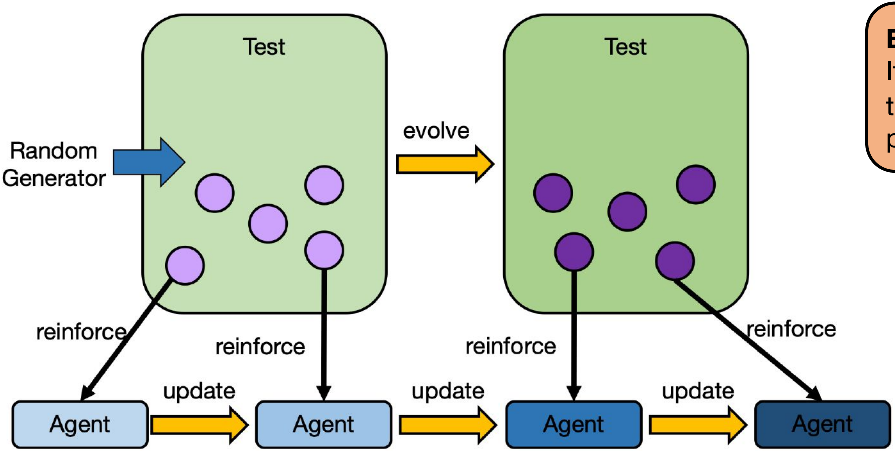
reign in?

a proposed paradigm

Gabor et al.
The Scenario Coevolution Paradigm: Adaptive Quality Assurance for Adaptive Systems.
International Journal on Software Tools for Technology Transfer 22 (2020).

Dieselgate Effect
If an adaptive system is much more powerful than its test, then it may prefer tricking the test to solving its actual objective.

Exam Effect
If an adaptive system can pass the hard tests, it can probably pass the easy tests.



Learning algorithms have inherent goals.

reign in?

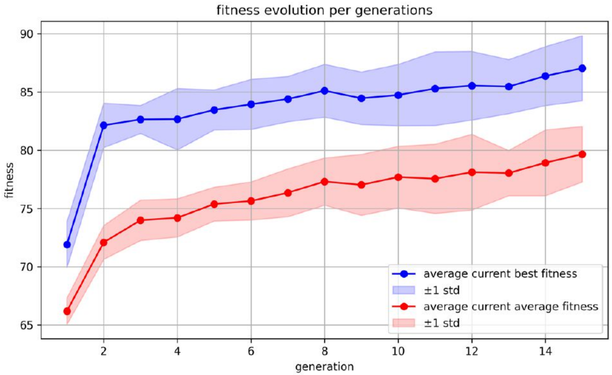
utilize?

work in progress

not yet published.

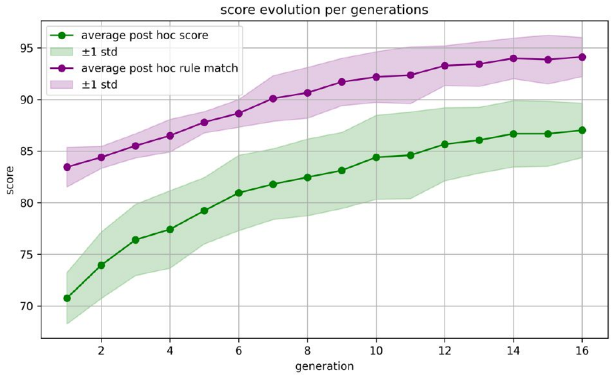
LLM

evolves an
animal description
and evaluation criteria for
animal descriptions
intertwined



increasingly better animal descriptions

optimization works!

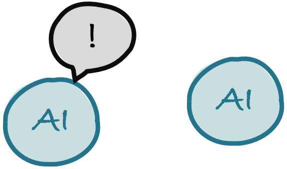


what the criteria say they want

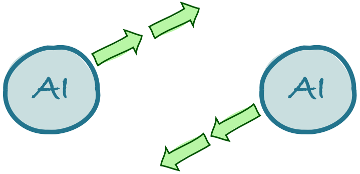
what animals are evolved

Complexity Evolution of AI-to-AI Interaction

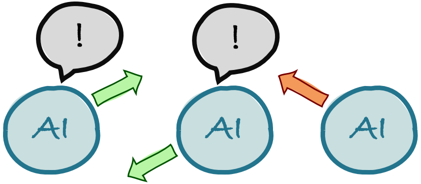
meeting in the field



converging interactions

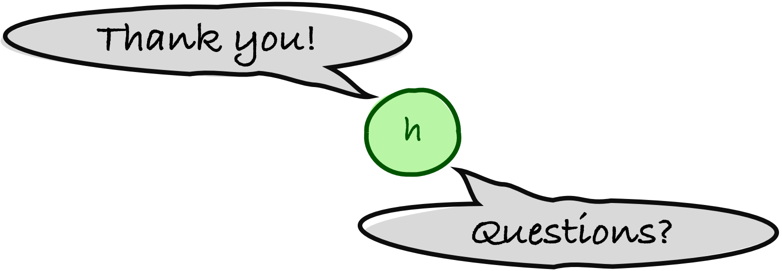


self-reinforcing behavior



What will AI ! do on its own?

Intrinsic and Emergent Goals in Optimization and Artificial Intelligence



Thomas Gabor
LMU Munich
I4CS Keynote, 2025-06-11, Munich