

This talk was written by a human.

Queens College, Oxford 18th Jan 2023 | Carla Zoe Cremer, DPhil

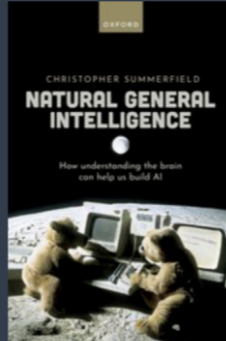


I was invited to ponder about what we learn about ourselves from studying artificial intelligence.

A pondering is what will follow. I am - as any semi-intelligent agent should do - still learning. These lessons will be personal: I cannot tell you about what we learn from AI. I can only tell you about a subset of things that I have come to learn.

I use talks like this to develop and try out ideas like I try out a new outfit, so thank you for this opportunity to sit down and write this little essay. If I make claims you feel are too strong but nevertheless worth refuting, I will have succeeded.

<https://xcorr.net/>



Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. Building machines that learn and think like people. *Behavioral and Brain Sciences*. 2017;40:e253. doi:10.1017/S0140525X16001837

The question subtly assumes an analogy, a similarity in process between human cognition and deep learning. This is neither false nor correct, but a question of the level of abstraction at which we are happy to accept an equivalence relation.

This argument is ongoing and richly populated with opinionated publications. It is also rather irrelevant for the question that concerns us today. Of course it is possible for us to learn about ourselves through AI even if there wasn't any overlap between our cognition and the most fashionable algorithms of our time.

In this short talk I want to contend that we do not learn about ourselves by the analogy of process, but via the inherently political nature of engineering intelligence.

You will encounter many more talks in which people tell you what intelligence is. I want to make you suspicious of it and attune you to the element of *judgement* that is fused into the concept. I want to

cast doubt on the air of respect that intelligence is accredited and propose that the process of building AI helps us primarily learn about what we value and whose values rule.

Lest I deprive myself of this opportunity to preach to you what intelligence is: the most commonly used definition is that it is the cognitive machinery that allows an agent to reach a goal in a complex environment.

The human brain is considered the most intelligent processing unit to date, but it seems that, and I quote:

"A brain is not a single thing [...] Our understanding will inevitably be fragmented and composed of different explanations for different parts"
(p.371, Cobb 2021)

Leg & Hutter , 2007
<https://arxiv.org/abs/0706.3639>

(The paper by Leg & Hutter lists 70 different definitions.)

TABLE I. LIMITATIONS OF DEEP LEARNING AS PERCEIVED AND NAMED BY EXPERTS FOUND IN [11]

Causal reasoning: the ability to detect and generalise from causal relations in data.	Common sense: having a set of background beliefs or assumptions which are useful across domains and tasks.
Meta-learning: the ability to learn how to best learn in each domain.	Architecture search: the ability to automatically choose the best architecture of a neural network for a task.
Hierarchical decomposition: the ability to decompose tasks and objects into smaller and hierarchical sub-components.	Cross-domain generalization: the ability to apply learning from one task or domain to another.
Representation: the ability to learn abstract representations of the environment for efficient learning and generalisation.	Variable binding: the ability to attach symbols to learned representations, enabling generalisation and re-use.
Disentanglement: the ability to understand the components and composition of observations, and recombine and recognise them in different contexts.	Analogical reasoning: the ability to detect abstract similarity across domains, enabling learning and generalisation.
Concept formation: the ability to formulate, manipulate and comprehend abstract concepts.	Object permanence: the ability to represent objects as consistently existing even when out of sight.
Grammar: the ability to construct and decompose sentences according to correct grammatical rules.	Reading comprehension: the ability to detect narratives, semantic context, themes and relations between characters in long texts or stories.
Mathematical reasoning: the ability to develop, identify and search mathematical proofs and follow logical deduction in reasoning.	Visual question answering: the ability to answer open-ended questions about the content and interpretation of an image.
Uncertainty estimation: the ability to represent and consider different types of uncertainty.	Positing unobservables: the ability to account for unobservable phenomena, particularly in representing and navigating environments.
Reinterpretation: the ability to partially re-categorise, re-assign or reinterpret data in light of new information without retraining from scratch.	Theorising and hypothesising: the ability to propose theories and testable hypotheses, understand the difference between theory and reality, and the impact of data on theories.

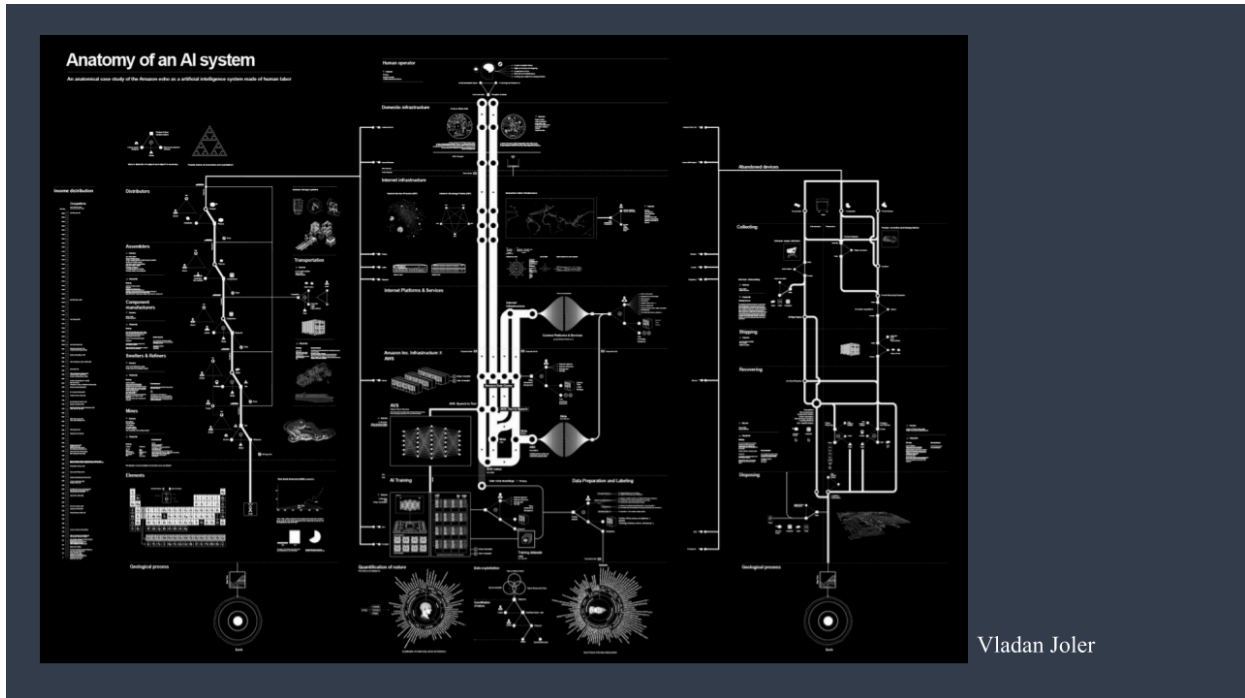
- Cremer, C. Z. (2021). Deep Limitations? Examining Expert Disagreement over Deep Learning. *Progress in Artificial Intelligence*, Springer.

Human intelligence includes the ability to decompose and recombine, make up rules, adapt preferences, button up shirts, reason with uncertainty, pick useful information, remember, and make shit up.

People vary tremendously in their emphasis across these faculties. I for one cannot simultaneously walk and marvel at Oxford without tripping and falling on my face. That can neither be excused as cute or clumsy, it's just a failure to achieve my goals.



Faculties expand and contract throughout a lifetime. My grandmother's Alzheimers has down-regulated her memory faculty, but up-regulated her humour and positively affected her dietary preferences, both of which are great signs of functional cognition.



The AI boom you witness is the process of how these cognitive faculties are being emulated, via **an ever intensifying process of stressed and lavish offerings of computer science graduates, gallium arsenide, burnout and sweat and even more numbers in even bigger spreadsheets,... offerings made by venture capitalists and prime ministers in hoodies, who hope to gain cosmic relevance, the admiration of young women and a long long life.**

The quest for artificial general intelligence makes it sounds like they are primarily trying to emulate the particular package of cognitive gadgets, i.e. the particular combination and integration of cognitive faculties that we might find in an average human, but what I think we see instead is the automation of optimisation.



Automation of optimisation

We see an algorithmically and economically feasible implementation of specific cognitive faculties or cognitive behaviours (language models) that are combined and implemented in ways to optimise screen time, trades, food deliveries, energy usage and to do so mostly by themselves. Demands will be met, utilities and profits will be gained, urges and pleasures be satisfied.

Why would we give up the degrees of freedom gained by having the mind's components like LEGO laid out in front of us? The result of our experiments will be alien intelligences, not one, but many.

I say this to emphasise the extent to which this process of engineering intelligence is *not* pulled along by some gravitational vortex of ground-truth at the centre of which Sam Altman will finally come to understand what intelligence was always meant to be.



(This of course is what you get when you ask OpenAI's DALL-E to generate Sam Altman holding a holy grail. Go figure.)

Instead, global engineering efforts and supply chains are pushed ahead by our current and peculiar notions of value, demand and ideology.

Sure you might say: I know that the automation of facial recognition, incarceration risk assessments or postal offices is politically controversial. I want to make the stronger case here that it's values and politics all the way down to the research.

You will know better than most that calling a person or a person's dog *intelligent* is more than a mere descriptor, a naming of a feature such as the dog's weight.

This dog weights 20 kilos.

This dog is intelligent.

Intelligent is an e-valuation, a judgement, and the difference in emotion we feel in reading those two sentences betrays the element of value that is inherent in the notion of intelligence.

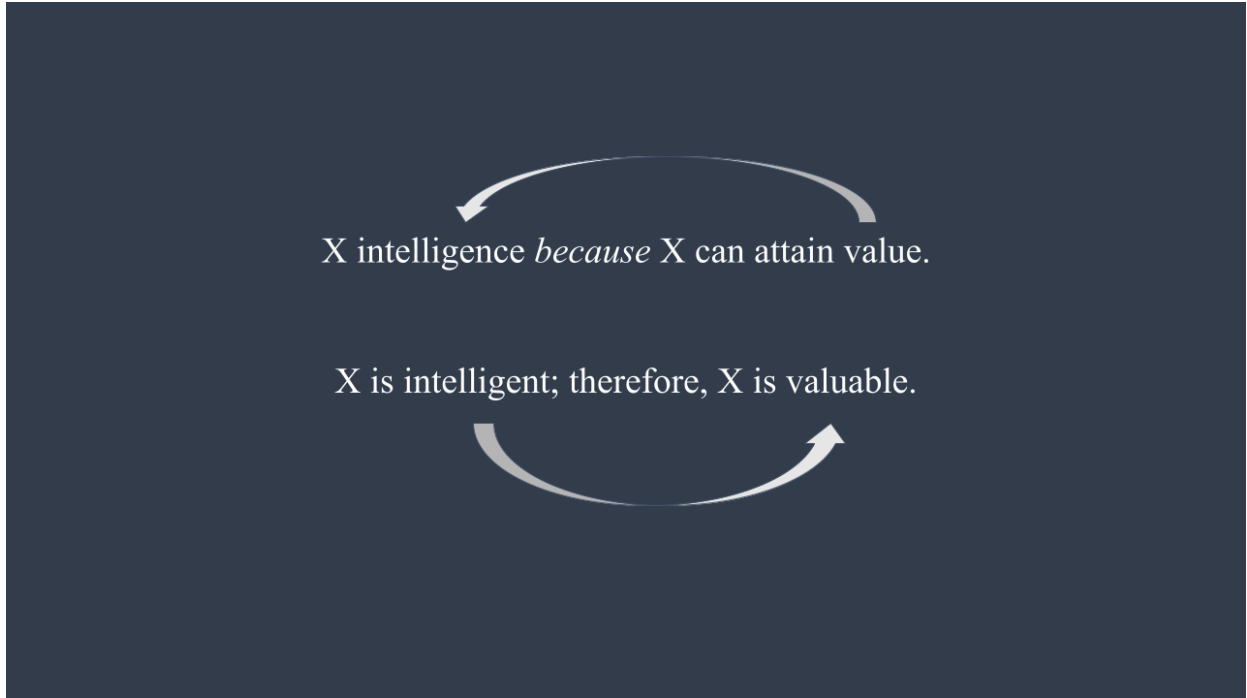
But our measures of intelligence and our benchmarks are in many ways measurements of whether an agent can attain something that we predetermine to be valuable.

A measure:

X is intelligent because X can attain value.

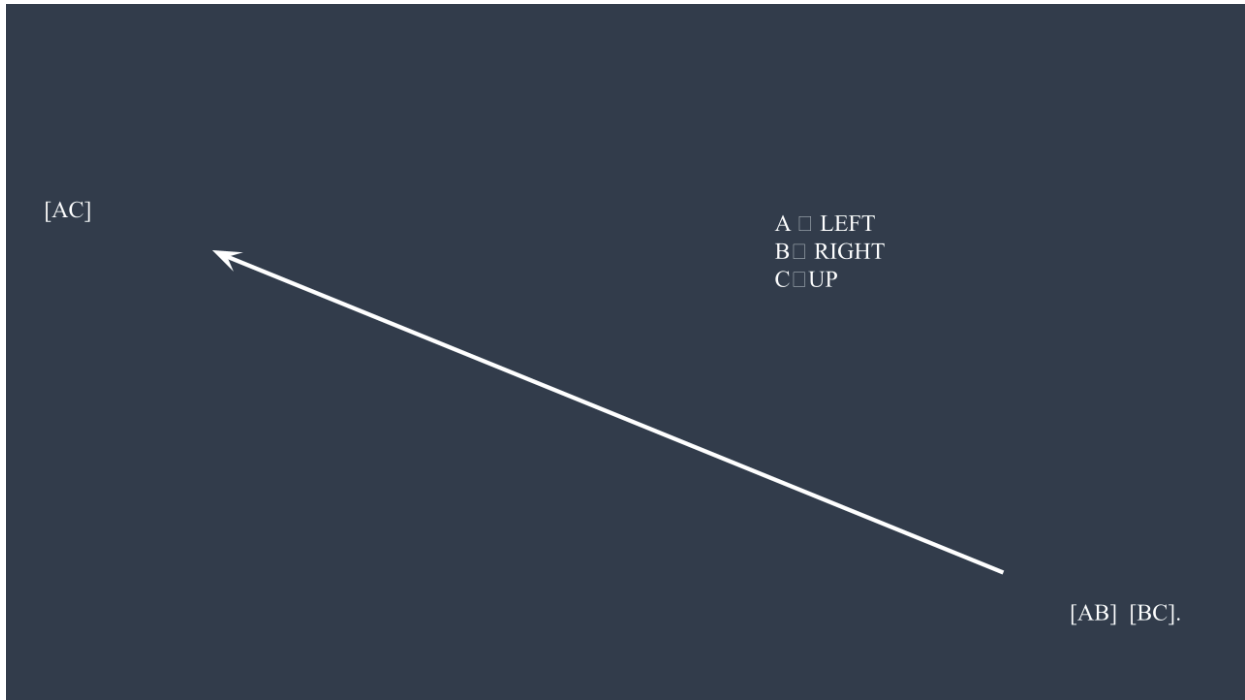
A compliment, an evaluation:

X is intelligent, therefore X is valuable.



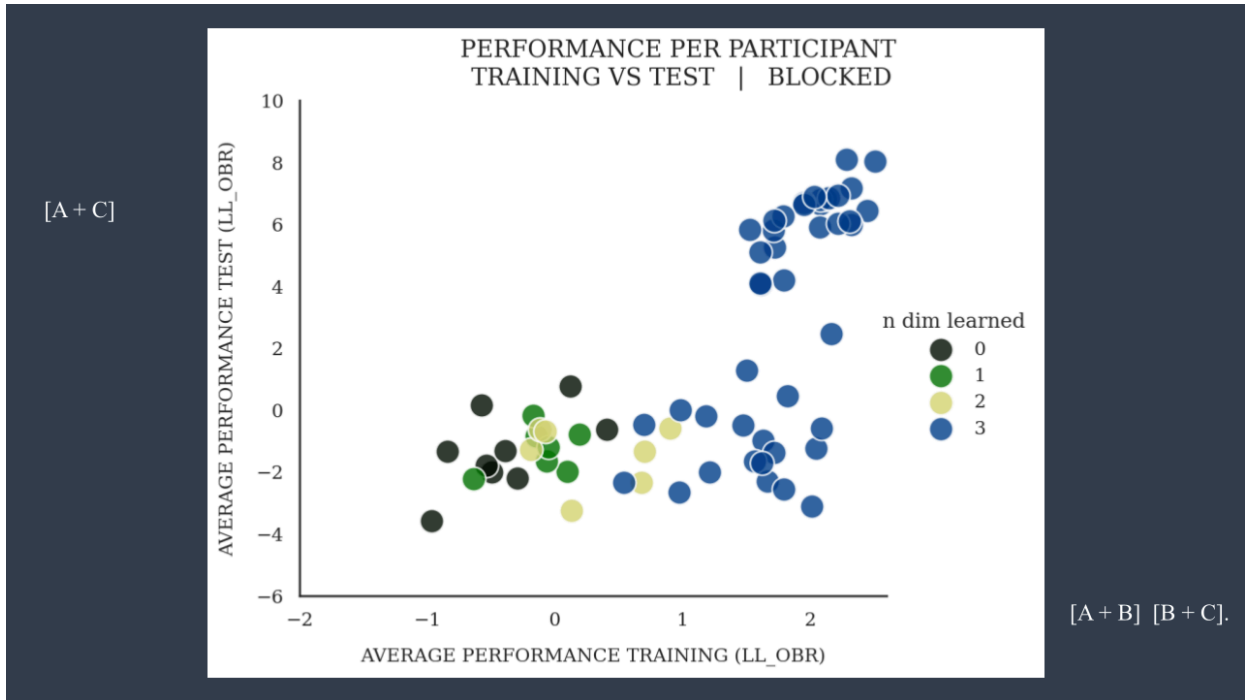
The circularity of the argument is apparent. Someone's going to have to fix the variable *value*.

Let's see how this looks like in my own research, in which I studied one cognitive faculty called compositional generalisation - a cornerstone of flexible reasoning. I study how people cut up the world and represent them in little pieces. This allows them to play LEGO with those bits of the world to solve new puzzles. Anyone who has tried to use GPT for the multiplication of large numbers will know that we haven't automated that process quite yet. I teach people to always respond in the same way to particular features. Let's say we have three inputs, A, B and C and three outputs LEFT, RIGHT and UP.



During a training phase, I make sure to only show them a few of all possible combinations, for example, participants see the combination AB and will be told they were correct if they press LEFT and RIGHT on a keyboard. There is a ground-truth, defined by my rules of the game as an experimenter.

But what we were interested in was a so-called generalisation phase, to see whether people learned a compositional rule, whereby each stimulus-response mapping was decomposed (into A, B and C) and independently recombined. In a generalisation phase I show people a new combination of input features: AC



That's what you see on the y-axis. The higher up a datapoint is positioned along the y-axis, the better this participant was performing during the generalisation phase. The further along the x-axis, the better they did during training. Basically, blue means the participant was good at the task.

Notice how I seamlessly ventured from a training context, in which the rules of success (A maps to LEFT) were arbitrary, but explicit, to a generalisation context in which the rules of success were arbitrary, but implicit. There is no ground-truth of how one ought to respond to a new input. It is the experimenter who defines the generalisation space.¹ We have presumed that it is adaptive to an agent to extrapolate like this.

This is of course not wholly arbitrary: we are all drowning in an academic literature to guide and certify our choices. Some ways of representing the world will be *better in respect to* our resource constraints and what we must be able to do as humans who survive in the world. There are bounds of persistence:

¹ Chris Summerfield really affected my thinking with that lesson.

an agent who repeatedly smashes into a wall will not be around for too long. Yet, **these bounds are wide and allow for much movement within.**

Look at all these (blue) participants (bottom right corner of my plot) who do perfectly well during training, but who completely fail to generalise to new combinations. In my study I call them failures or 0-D generalisers, but this is an online game and for all I know they generated wonderfully elaborate jazz dance choreographies in response to seeing AC! Or maybe they just thought that AC is not a linear combination of A + C but something non-linear, more rich and complicated. This isn't a failure of capacity, but of assumption, motivation or even taste.

The ability to decompose AB into A + B is a skill without which an AI will be fundamentally limited. But where and how and to what ends this ability will be applied is a matter of politics.

Whatever you study, please do not exempt yourself from having opinions on the matter. Why don't we test the limits of democracy on algorithmic design?

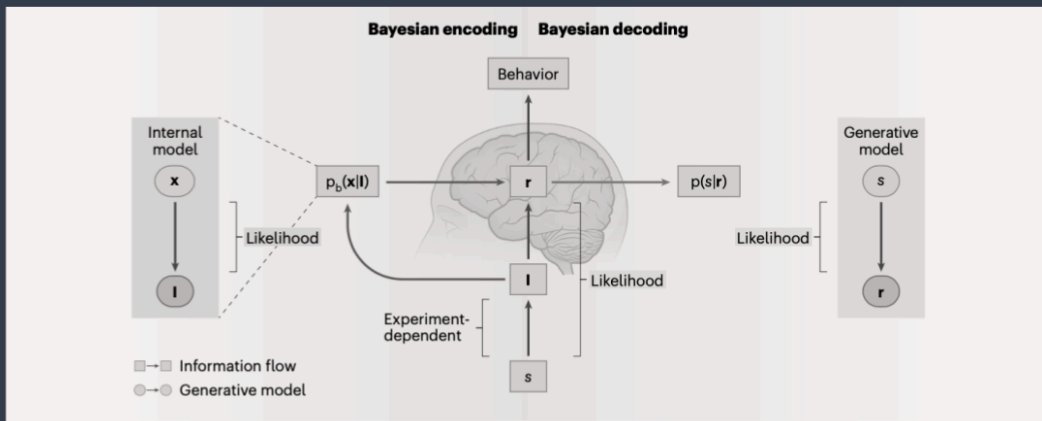
In this apparent 'age of AI' are we indeed getting better at describing and emulating these cognitive faculties? Undoubtedly so. Remember when we thought we were different, because if you perturbed an image of a panda the slightest bit, the AIs thought it was a bus but we still saw a panda?

'frog'



arXiv:2308.06887v2

Adversarial attacks are still an issue, but it turns out we can fool humans just the same and predictably so.



10.1038/s41593-023-01458-6

So what if we are input-output processors? Well, we already knew this, didn't we?

At some level of abstraction this was always a correct description.

What if we can describe the average human's inner machinery with greater precision, more detailed maths?

The wind that shakes the barley, 2006

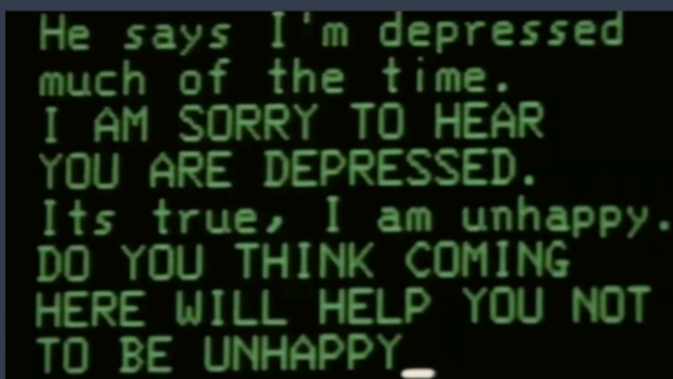
How does this change how you look at our lover?

If it does, was it really love?

For the remainder of this talk, I will assume that, within our lifetime, intelligence will become legible and cheap.

Intelligence will be cheap.

I don't mean the human package of intelligence, but **the automation of various cognitive faculties, assembled in new ways to solve problems with a subset of profitable solutions.**



He says I'm depressed
much of the time.
I AM SORRY TO HEAR
YOU ARE DEPRESSED.
Its true, I am unhappy.
DO YOU THINK COMING
HERE WILL HELP YOU NOT
TO BE UNHAPPY_

Our need to feel seen and have our needs met
is far greater than
our motivation to exercise discernment.

ELIZA

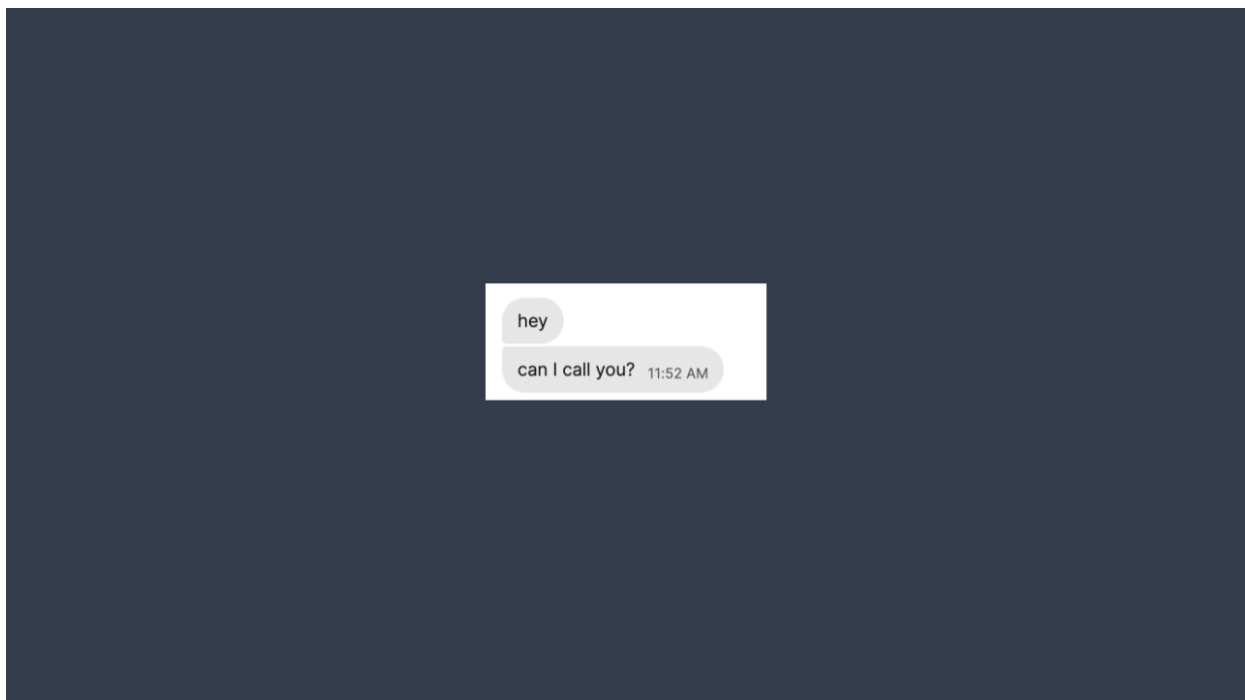
Role play with large language models

<https://doi.org/10.1038/s41586-023-06647-8> Murray Shanahan^{1,2,3,4}, Kyle McDonell^{1,2,3,4} & Laria Reynolds^{2,3,4}
Received: 10 July 2023

We have known since the chatbot experiments in the 60ies that there's little need to emulate real cognition or real attention. You can get much of the dangers and much of the benefits from AIs that pretend they can speak or pretend that they listen.

Our need to feel seen and have our needs met is far greater than our motivation to exercise discernment. Our focus and discernment is energy-intensive and we apply it sparingly, as any rational, resource-bound agent should.

Will cheap, available intelligence make us lose our interest in one-another? One lesson I have taken from watching AI tools disseminate into our social fabric is how much we actually *use* each other.



For warmth at night, to act out our traumas, as guard rails for the moments when our own prefrontal cortex feels off balance and we lack cognitive control to regulate our emotions or as advisors for when we are willing to slide back into a relationship that wasn't good for our health.

And we like being used, being useful to our people.

As AI advances, there will be less of a need to go to other people for utility. GPT assist will talk me down from a high horse or the brink of a break-up and play me Jack Wilkins and unfollow Jockstrap on Spotify.

But cries that warn we might lose the practice of getting utility out of other human bodies and minds seem all too pessimistic about the nature of our relations. Call me an irresponsible romantic, but utility monsters are not all that we are and I for one look forward to a world in which we choose each other more freely than we might do now - not because we need each other, but because we want each other.

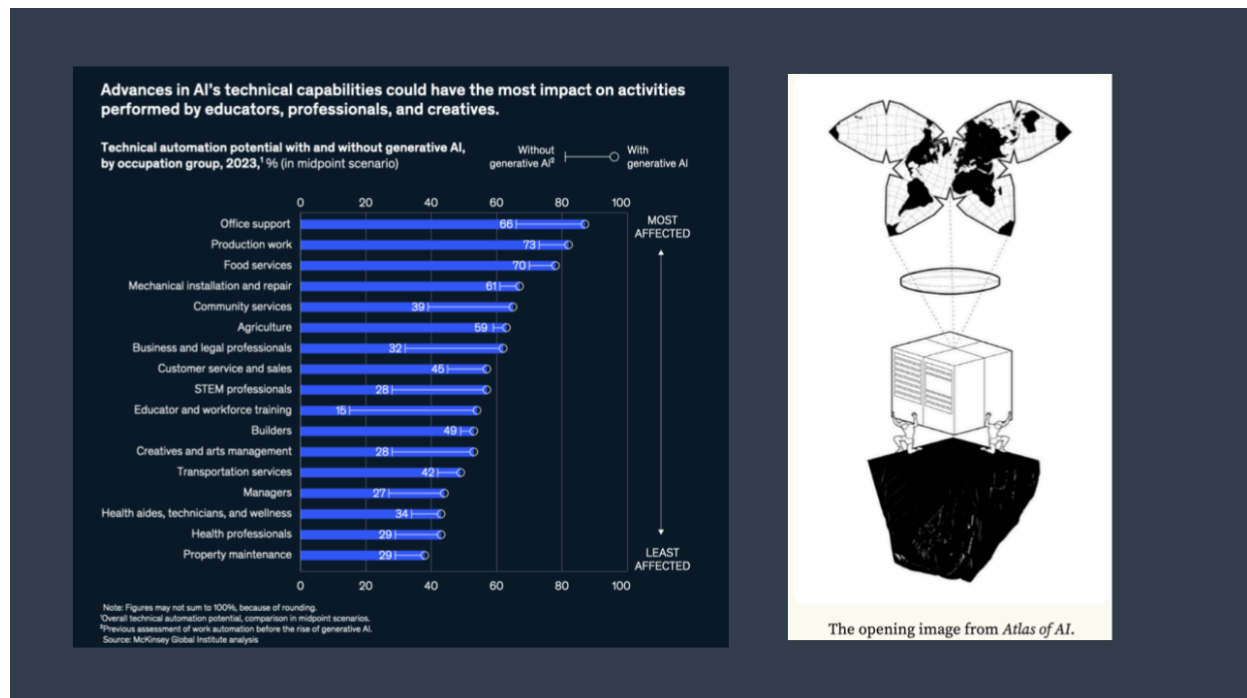


Someone real paying attention to you will be the ultimate compliment.

The price drop on the product *intelligence* may eventually come as a shock to those whose identities have been intertwined with being especially smart snowflakes.

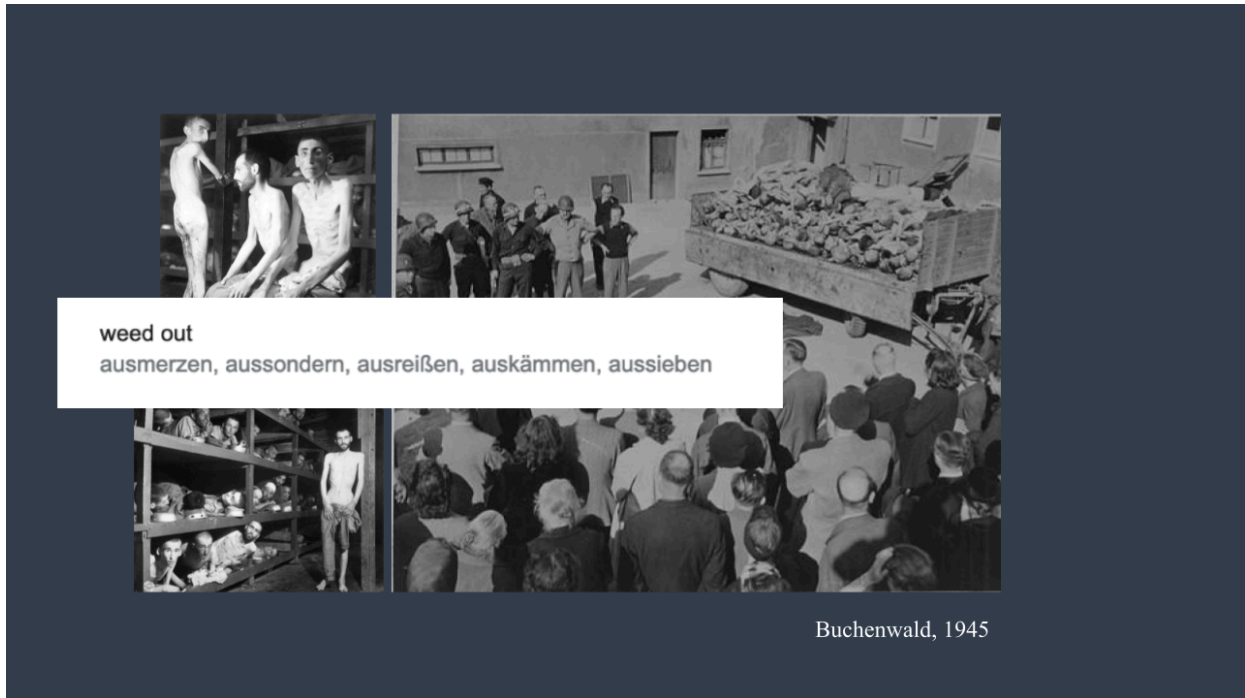
You are here because you exceeded some threshold in a system of benchmarks and exams that branded you as intelligent. Some wiser part of me would hurry us all to start digging for new origins of our self-esteem.

In fact, **whole nations will need to come to believe that our capacities to beat a benchmark is not at the core of what we adore in one another.**



If we continue to place moral value on cognitive capacity, **we will find our individual lives in a race with whole industries that optimise alien intelligences.** We cannot win such a fight. **We can amend the rules of the game so that the winner works for us.**

The moral devaluation of intelligence is a necessity. Because history tells us what happens under such moral experiments: when performative attributes and features become the explicit proxy for an otherwise essentialist notion of otherwise ineffable value, whole economies can be redirected to optimise, to weed out features that are supposedly anti-correlated with said value.



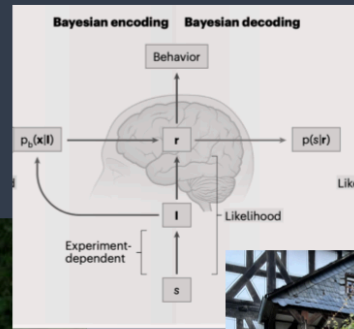
The German language appears to have a disgusting amount of synonyms to describe *weeding out*. What follows is monstrosity, monotonicity, but surely, certainly not intelligence...

Now I too just used *intelligence* to cast judgement. Didn't we start by saying intelligence is merely the ability of an agent to achieve goals? Did I mean to say wisdom? **Or did I really use the term intelligence as a placeholder for my politics, my latent Christianity?**

cling on to and develop the particularities of your own alien cognition



The wind that shakes the barley, 2006



my granny, 2023

I must confess that the more I look at intelligence, the less interested I become. I still marvel at the processes, but a standardised measure of processing capacity turns out not to have much in common with what I find intriguing in the idea of a person.

When they next tell you what intelligence is, go look for their values, look for their votes.

Resist the standard, cling on to and develop the particularities of your own alien cognition. We will need them in the future that we will inhabit.

Thank you.