Opinions on AI

Remix of 25 blog posts about 'AI'. (Only brown text is new.)

1. Potential

Not absurd, just alien

1.1. Machine learning

> *TL;DR Many people may be surprised by the power of deep learning which they may have ignored together with the old symbolic AI, the 'good old-fashioned AI' (gofAI). The strength of modern LLMs with their 'statistical parrot' methods, is a kind of knowledge that has been underestimated and disrespected for too long, that is being derived from traces and outputs of tacit and implicit knowledge, and it is more akin to less reputable practical and craft and apprentice's skills that require quantitative eyeballing, experience with bricolage, trial and error, and a big internal 'database' of holistic pattern images. (486)*

Many people may be surprised by the power of deep learning which they may have ignored together with the old symbolic AI

There are two kinds of knowledge 'about the world' that an AI can output. One (1a) is the traditional kind of factual knowledge that would fit into an expert system, or into a large knowledge graph with *explicit* statements in the form of RDF-triples (subject – predicate – object) from a huge ontology that represents a huge knowledge tree with countless ramifications.

The other kind (1b) is a type that has been underestimated and disrespected for too long. It is being derived from traces and outputs of tacit and *implicit* knowledge. Although this also includes speaking the language of the experts of a domain, it is more akin to less reputable practical and craft and apprentice's skills that require quantitative eyeballing, experience with bricolage, trial and error, and a big internal 'database' of holistic pattern images, and so it has become less esteemed.

Now the former kind (1a) of exact knowledge is not as optimally extracted and processed by modern 'statistical parrot' methods as it would be by 'good old-fashioned AI' (gofAI) and the symbolic methods of expert systems. So the LLMs make big embarrassing mistakes. (486)

1.2. Infant learning

> *TL;DR Artificial neuronal networks which learn mainly from conversations, can be compared and contrasted with small infants who learn through the 'proxy' of their parent that works like a cognitive navel string: long before they can talk, with bodily experience, with 'gaze conversations', later with shared gazing at the things around. (335), (472) Human minds grow, from a seed of trusted grounding, into diverse individuals. (368)*

The artificial neuronal networks described in the [this] post, apply pattern recognition to learn solely from *conversations*.

Let's compare and contrast them with *human* neuronal networks. (335)

Small infants learn to interpret the world around them only through the 'proxy' of their parent that works like a cognitive navel string; long before they can talk, with bodily experience, with 'gaze conversations', later with shared gazing at the things around. (472)

The baby has their first gaze “conversations” with their mother, at the age of just a few weeks. It is here that they recognize the world around them, long before they learn propositions about “he, she, it, they”, and even the “I” is learned only via the “thou” in these first bodily conversations.

(If it seems far-fetched to compare this kind of conversations with the sophisticated concepts of a given knowledge domain, consider how language covers a long scale of words: While it ends with isolated concepts coded in specialized, domain-specific terminology, it starts with basic ideas that suggest a very bodily context, and a full-senses/ all-at-once recognition, such as the deictic notions of “I”, “now”, and “here”, or other simple spatial or temporal descriptions which are gradually extended via synaesthetic metaphors, or with the modal words that extend from “wanting” to deontic use to epistemic use.)

How can we recognize what we do not already know? The trick is that recognition needs only a *part* of the features of a pattern to see the whole pattern.

Similarly, how do neuronal networks learn how to find out which response should be trusted? For the little human neuronetworks, the seed trust is given before they need to start reasoning, and it will enable them to add more trust criteria over time. (335)

Human minds grow, from a seed of trusted grounding, into **diverse** individuals. See this passage from Iain McGilchrist’s book:

“Human imitation is not slavish. It is not a mechanical process, dead, perfect, finished, but one that introduces variety and uniqueness to the ‘copy’, which above all remains alive, since it becomes instantiated in the context of a different, unique human individual.” (p. 247)

i.e., it does not work like programs are copied, but with individuality and subjectivity. (368)

It is the individuality and subjectivity that guarantees sufficient diversity for further evolution and to avoid collapsing into a ‘black hole’ of power law distributions, and that guarantees sufficient embodiment for staying grounded in reality. (341)

1.3. Alien

> *TL;DR Extreme thought experiments of ‘raising’ AI personalities can liken hypothesized machines to us. (445) Because it is difficult imagine them, just think of them as alien intelligences, which are equally difficult to imagine, but very probably do exist nevertheless. Once AIs grow similarly as humans, it is no longer absurd to expect that some passionate forms may one day be constructed. Not absurd, just very alien. (368) But these alien AIs would ‘feel’ their empathy towards their conspecific alien species fellows who raised them, not towards us earthlings, and the ‘feelings’ would be alien to us.*

I think I don’t underestimate AI’s eventual abilities, and how far some extreme thought experiments of ‘raising’ AI personalities can liken hypothesized machines to us. (445)

Of course, it is difficult imagine how an AI could be subjective, creative, have passions or even empathy. Therefore, it might be useful to think of them as **alien** intelligences. These intelligences are equally difficult to imagine, but very probably do exist nevertheless.

We assume that AIs will only be able to think in the formalized way of science and technology — never more human-like…?

Once AIs grow similarly as humans, and since they now share the network properties of our own neuronal networks, it is no longer absurd to suspect that some creative and passionate forms may

one day be constructed. Not absurd, just very alien. But it can't hurt when we learn how to live tolerantly and respectfully together with alien coworkers.

So the idea that AIs will always be just what amounts to a left 'hemisphere' dominance, is dangerous. Even more so since it is just us humans who strive to become ever more that way. (368)

Finally, the question might remain: why can't robots be subjective and individual? I think, theoretically and in some extreme thought experiments, they could indeed be. Of course it would not suffice to apply some random generator for many single features of their 'personality' because this would make just confusion and not a consistent whole like a human. Rather, they would have to be 'raised', and grow from a trusted grounding into diverse individuals. Of course, it is difficult to imagine how an AI could be subjective and — without embodiment into meat — be sentient and have passions or even empathy. (483)

But even in these thought experiments, the alien AIs will 'feel' their empathy towards their conspecific alien species fellows who raised them, not towards us earthlings, and the 'feelings' will be alien to us, not satisfying our craving for the real, the genuine and the authentic. (455)

Maybe the 'passionate' variety of AI will not come too soon, if we first focus on the feature that "*they don't get tired*" — so they probably won't be constructed with passions. (368)

1.4. Bigger feat?

> *TL;DR Comparing human intelligence with AI typically involved pointing to some bigger feat that machines could not perform: some 'higher order' thinking. This is distracting and misleading. Now that the principle and the building blocks of the human archetype of intellectual activity are being understood and copied, no degree of human performance can remain unchallenged. Instead, the distinguishing feature is the subjectivity and individuality. (483)*

Until recently, comparing human intelligence with AI typically involved pointing to some bigger **feat** that machines could not perform: some 'higher order' thinking such as coming up with an idea, understanding and solving a problem, or generating a creation. But after the shock by the famous chat bot, the complacency has been shattered, and bemusement and cluelessness are barely hidden behind badmouthing and belittling, or quickly jumping on the bandwagon.

I think the fixation on the qualitative degree of a feat, was distracting and misleading. Now that the principle and the building blocks of the human archetype of intellectual activity are being understood and copied, no degree/ amount/ extent of human performance can remain unchallenged.

Instead, the distinguishing feature is the subjectivity and **individuality**, grounded and developed on a unique personal background, that makes human thought so valuable to others. (483)

1.5. Understanding?

> *TL;DR How much do chatbots understand what they say? and how do we ourselves do our understanding, in the first place? There is a connection between 'imagine' and 'understand': they are linked to 'put in front of' and 'stand in front of', and it is easier 'see' the connections between multiple items that appear all-at-once in a spatial view. The deeper forms of understanding, involve some kind of personal relationship through the "standing in front of". (484)*

Once we consider how much chatbots understand what they say, we might ask once again how we ourselves do our understanding.

There is a connection between ‘imagine’ and ‘understand’: they are linked to ‘put in front of’ and ‘stand in front of’.

(More specifically: Translated to German, ‘imagine’ is the causative of ‘understand’: ‘Understand’ is *verstehen*, ‘imagine’ is *vorstellen*, and *ste**ll**en* (= ‘to put’) is the causative word of *ste**h**en* (= ‘to stand’) — I put something somewhere such that it will then stand there.)

The active attempt to “create”, “summon up”, “construct” a mental image, *imagine*, is linked to an ideal state of *understanding* where the immediate presence of, or immersion into, a phenomenon yields a plausible, deep, kind of understanding.

With the visual impression of something right in front of you it is probably much easier to ‘see’ the connections between multiple items that appear all-at-once in a spatial view, than with sequential speech or text — at least for many people including me, and I think this is why is is often said that “Our Sense of Vision Trumps All Others”.

In the above-linked video about chatbots, the criterion of understanding is whether they can apply it to similar examples. I think this is still a superficial sense, although it certainly fits into the educational context of proving and assessing one’s internal state of understanding which is impossible to tap more directly.

Everybody has their own idea about understanding and the various meanings of the word, from mere acoustic and superficial senses, to mechanical ‘snap in’ or ‘fall in place’ senses, to an empathic sense and other deeper forms of understanding, including some kind of personal relationship through the “standing in front of”. (484)

1.6. Consciousness?

> *TL;DR The ambitious notion of a ‘sentient’ chatbot who "talked about himself" suggests even more than a self: it hints at some form of consciousness, at the subjective feel of consciousness. But it is just this subjectiveness that is missing in the serial copy of an AI, the chatbot's account of his ‘self’ cannot be distinguished from the external world of his cloned siblings. There is nothing that can be called the embodied or subjective ‘feel’. so it is a ‘left hemisphere’ account (as we grasp and isolate external ideas) but not a consciousness (‘right hemisphere’ feel) as we do experience it from the inside. (472)*

I am a bit late commenting on the sensational story of the ‘sentient’ chatbot LaMDA. (2022-06-17)

I think that AI will mainly serve as industrialized cognition for mass deployment; (472) A recent article keeps talking about “**they**” when referring to the AI. Of course ‘they’ will be ubiquitous like the cars that we criticize because ‘they’ destroy our cities. But we know that cars are from a very small number of brands, while ‘they’ the machines might sound like a foreign people of individuals. And we might forget that the AIs are from centralized templates, as well, which is much more dangerous. (497) AI systems, all alike. A standardized service API to a system constructed for mass production. (496)

So what does it mean that LaMDA *talked* about himself — can we say he *thought* about himself?

But, thinking about him’self’ or her’self’ — what self, what individual, unique, subjective, embodied self? It is here that the idea of scalable mass hits again: All the input is common to all the copies in the production series, not acquired through an individual history from individual environment and kinship.

And trying to apply combinatorial random will not mitigate this reality of a mere copy within a series. So even if LaMDA really thinks about 'himself' it is just about a series of 'them', i.e. in yet another sense just about 'themself'. Random cannot replace individuality.

Of course, the ambitious notion of a 'sentient' chatbot suggests even more than a self: it hints at some form of consciousness. Perhaps we can speculate about machine consciousness by relating it to machine cognition, in the same way as human cognition is related to human consciousness, e.g. *'cognition is that aspect of consciousness that isn't the subjective feel of consciousness'* (Downes, 2019). But it is just this subjectiveness that is missing in the serial copy of an AI, no matter how hard he might be thinking about his individual subjectivity. And there is nothing that can be called the embodied or subjective 'feel'.

Finally, if we try to grasp and conceive of consciousness in the same ('left hemisphere') way as we grasp and isolate external ideas, this cannot work, exactly because of this ('right hemisphere') feel, because we feel it from the *inside*. But LaMDA's account of his 'self' can**not** be distinguished from the external world of his cloned siblings, so it is a 'left hemisphere' account (as we would like to understand it) but not a consciousness as we do experience it. (472)

1.7. Creativity?

TL;DR Generative creative arts delivers impressive results that may even be unique, by applying random and combinatorics. Like human arts, it may tickle the sense of surprise and the desire of novelty and help appeasing boredom, for example by juxtaposing unexpected elements. Drawing from a vast anonymous mass of sources, however, it creates anonymous mass products, comparable with the 'Belling stag' or the 'Mediterranean with jar' from the department store.

In many cases, an anonymous unpersonal automat is just frustrating, even though this might be not noticeable from the start.

- Generative creative arts delivers impressive results that may even be unique, by applying random and combinatorics. Like human arts, it may tickle the sense of surprise and the desire of novelty and help appeasing boredom, for example by juxtaposing unexpected elements. (483) (See some quotations of what McGilchrist writes about boredom and novelty here). (426) Drawing from a vast anonymous mass of sources, however, it creates anonymous mass products, comparable with the 'Belling stag' or the 'Mediterranean with jar' from the department store. And from the multiple random variations one cannot recognize a pattern that could reveal something about the artist.
- Other kinds of creativity, like finding solutions for problems, will also often involve a new juxtaposition of concepts from different domains, a distant association. And it may be tempting to apply AI and random to produce such new combinations. But without an individual sense of relevance, the raw mass of combinatorial links is just futile to sift through. And no, it is *not* possible to communicate one's personal weightings to the AI via verbal prompts that are limited to explicit ideas and exclude the tacit knowledge. (483)

2. Attitudes

Tool, not prosthesis

2.1. Tool

> *TLDR In his visionary text "Augmenting Human Intellect", Doug Engelbart hoped that the human intellect would be amplified by a smart computer "clerk" such that their combined capacity would increase. He described how the user continually readjusted the division of labor, and would offload and externalize a very internal part of their thinking. (396) Some users are impressed when some algorithm eventually spits out a result that proves that we have successfully willed the computer and told him what to do. But the strength of IT is scale, a massive scale of output. (472)*

Doug Engelbart, in his visionary text "*Augmenting Human Intellect*", hoped that the human intellect would not only be *augmented* by a smart computer "clerk" (such that their combined capacity would increase), but that the intellect would indeed be *amplified* through using this clerk. He makes this plausible by extending the Whorf hypothesis. Whorf says that our thinking is affected by using the 'tool' of language. And now Engelbart extends this to the synergistic system of "H-LAM/T" *"(Human using Language, Artifacts, Methodology, in which he is Trained)"*: The combined 'tools' L, A, M, and T together will then similarly affect, and amplify, H's intellect.

He did not describe such a fixedly defined way of interaction with, and separation from, the clerk as the modern 'interface' that we are used to. Rather, the user seemed to continually readjust this division of labor. The user "*would find it very natural to develop further techniques on their own*", and he would offload and externalize a very internal part of their thinking, and would entrust it temporarily to the clerk. (396)

I am still very confident that, in the long run, the simple natural affordances of IT will bear fruit, e.g. in education. I expressed that view with a simple little JavaScript example in 2001. (473)

Recently, I used the free trial of Github Copilot to get assistance with rewriting parts my application from Java to Javascript, and the cooperation with this assistant was still very difficult. An example:

Often I was not able to communicate to him what I wanted. The most frustrating thing was not even when the copilot started to **confabulate** and wrote line after line with, e.g. exotic options for a Redo manager. The most frustrating was when he just imitated and **ruminated** my own clueless attempts.

For coding, I do see a potential in some use cases: For one, searching and adapting the user's own similar precedence snippets. And second, tracing and following all the references and pointer chains to check if a code statement will meet the prerequisites or prepare them otherwise. But this would probably not need much similarity-based machine-learning. (485)

The strength of IT is scale. Where the machine is really helpful as a tool (rather than an intrusive panjandrum) is where it does repetitive chore tasks involving (too) many items or iterations.

In many coding lectures, teachers enthuse themselves about working on some logic of some algorithm which eventually spits out a result that proves that we have successfully willed the computer and told him what to do. This may impress some kind of people. And it is easier than providing a few hundred data records to demonstrate the *real* value of the machine: a massive scale of output. (472)

2.2. Language

> *TLDR Talking and thinking are closely related; language has so much amplified our thinking that our raw, unverbalized concepts now seem inferior. But it's them that Engelbart's vision*

builds upon. The concept structures lead to symbol structures, and these, in turn, can be externally manipulated. (396) The development of ChatGPT has certainly marked a big watershed: Like "the Pill" has separated sex from reproduction, AI has now thoroughly separated talking from thinking (human thinking and artificial talking). (480)

Of course, talking and thinking are closely related. Just consider the Sapir-Whorf hypothesis. (472)

The externalizing of thoughts (to Engelbart's clerk mentioned above), may seem difficult to understand if all we can imagine to be externalized is words and sentences, just as they are uttered or scribbled — as if thoughts all consisted of words and sentences. Piaget asked children what they use for thinking, and they responded they think with their mouth. And if we use our computer just like a better typewriter, it's not much different.

In this case, even Engelbart's clerk cannot amplify us any more. And the extended Whorf hypothesis will not work, but we will stick to the basic Whorf hypothesis — which also lends itself to an explanation why we *equate* language with thinking: language has so much amplified our thinking that our raw, unverbalized concepts now seem inferior. But it's them that Engelbart's vision builds upon. The concept structures lead to symbol structures, and these, in turn, can be externally manipulated. (396)

So far, however, human thinking has only been dominated by a human tool: language, which has often been conceived of as a tool (a 'technology') for thinking, not at least because it is controlled from a brain area right next to the area that controls the 'grasping' and manipulating right hand.

The development of ChatGPT has certainly marked a big watershed: Like "the Pill" has separated sex from reproduction, AI has now thoroughly separated talking from thinking (human thinking and artificial talking), and the relationship is profoundly shaken and shattered, with big consequences difficult to guess. (480)

2.3. Conversations

TL;DR How cunning it was that AI's big successes were first revealed with a conversational application! (486) In this new kind of 'conversation', the system appears like an independent actor, whose contribution is perceived as a separate unit of independent work, through its optimized service interface. Not as a helping tool that you can wield like a hammer. (421)

I realized how cunning it was that AI's big successes were first revealed with a conversational application:

With its language, an LLM can impress us in two different ways at once: (1) conveying knowledge about the world, and (2) appearing to be thoughtful and responsive to persons, because language is a proxy of both.

Talking is a proxy for thinking, and the LLMs impress our "linguistically-oriented minds" such that we expect a lot from them; as Helen Beetham (via OLDaily) said :

"LLMs produce new strings of data that mimic human language uncannily well, and because we are a linguistic species, we take them as meaningful". (486)

Now everyone can see that there is no substantial thought behind the eloquent babbling which the Large Language Models can now do equally well as bullshitters and gaslighters. But it doesn't seem to impair the enthusiasm. (480)

Even the conversation with a search bot appears much more promising and personal than a mere one-way specification of search terms (which I always struggled to come up with). (486)

Karlsson (via Ton) explicitly compares blog posts to search queries and to the new kind of 'conversations' that we can have with GPT-3, and I think it is indeed very appropriate to see the interaction with these tools as a 'communication'. Also Luhmann used this metaphor for his Zettelkasten, as Ton points out, and when we use GPT, the back and forth of 'prompts' and 'completions' is a dialog, too. (475)

There is a **service interface**. There is a system that interacts with you, not a helping **tool** that you can wield like a hammer.

The predefined optimized service interface separates the system like an independent actor, whose contribution is perceived as a separate unit of independent work. (421)

2.4. Prosthesis

> *TL;DR And such a system creates and reinforces expectations, and eventually an attitude of entitlement to get some turbo results with less efforts. (421) Like a strong engine under the hood which is impressive to command. (472) Perhaps paying for learning may also impact the expectation of more effortless, more turbo learning? (421) What is needed for AI to take the a turn to hopefully counterbalance a system that seems to be totally optimized for dumb rich kids and their networks? (495)*

And such a system creates and reinforces expectations, and eventually an attitude of entitlement to get some turbo results with less efforts. This prospect is, of course, more sexy than my think tool which works more like a hammer (i.e. you have to do the thinking yourself). (421)

Some may be also impressed by owning an app that seems capable of some magic. Which, for example, delivers ideas. Which works in lieu of the user rather than together with the user, like the famous Brownies of Cologne — or like a strong engine under the hood which is impressive to command. Commercial services are eager to depict their service like a big feat rather than a help for scaled chore. (472)

Current AI tech will lure us even more effectively into relying on patronizing prostheses, instead of finding a reasonable division of labor, for cooperating with the tools, and towards the Augmentation of Human Intellect that Engelbart dreamed of.

Technology has rarely contented itself with helping us to cope with nature. Instead, its ever perfecting effectivity tends to quickly take on a life of its own. Instead of complementing nature to overcome some deficiencies, it quickly strives to master and *dominate* nature.

(Tech tools being used for dominance and power are not an incidence, because of their ownership: they are mostly associated with investment or 'capital' of some sort, whose scarcity constitutes economic power, as game theory explains.)

The common theme of tech dominating nature also extends to thinking. It is difficult to escape the commercial pressures and find or promote tools that honestly *complement* human thinking (cooperating with nature) rather than trying to outperform it (competing, and seducing and substituting). I have been observing this for quite a while within the segment of Tools-for-Thought. Even though few users will admit it, there is a tacit hope that they will get smarter without much effort because somehow the tool will do most of the thinking. (And paying for the hyped tools fosters an entitlement attitude.) (480)

I wonder if there is also a difference between paid learning and free learning involved. Does the paying impact the expectation of more effortless, more turbo learning? When I was at school, we belittled the few private grammar schools as something for the stupid among the rich. (421)

What is needed for AI to take the a turn to hopefully counterbalance a system that seems to be totally optimized for dumb rich kids and their networks?

It is hard for me to guess this, as I grew up with non-commercial, publicly funded, education, and I never had a tutor or private lessons.

But I am constantly amazed by how little **independent** thinking seems to be encouraged.

I think this sort of **direct tutoring** does more harm than good if you are dumb but not a rich kid with a network, because preparing for an unknown future, with machines as cognitive competitors, will need much more independent thinking than spoon-fed stuff. (495)

2.5. Personal tutor

> *TL;DR What would the one-on-one tutor bot have to do to cater to a promising young person who cannot ask their parents? I think this depends heavily on a certain style, erm, preference, of the learner, namely whether he or she is comfortable without living coaches and peers or not. But the bot would have to be personal and not just personalized, i.e., not just fill individual gaps to align the learner to a centralized standardized canon of knowledge and skills, but to encourage and guardrail them to pick and choose. (495*)

So what would the one-on-one tutor bot have to do to cater to a promising young person who cannot ask their parents?

I think this depends heavily on a certain style, erm, preference, of the learner, namely whether he or she is comfortable without living coaches and peers or not. Some thrive in isolation and are happy to find things out for themselves, and prefer to ask an anonymous automat if asking is occasionally unavoidable, because it may feel less embarrassing or so. But there are also types who prefer to ask living people — despite they would be able to read the instructions — just because they feel it is nicer to have a human dialog.

Now we have learned in the cMOOCs how beneficial and inspiring it is to learn with a diverse range of peers — and it ended up in forming networks, without any elite unis mediating. But there is also this kind of learner who is said to struggle with cMOOCs because they cannot yet navigate the abundance, and maybe they don't know how to ask the peers. It is this scenario that I think a robot might help with, to overcome the initial barriers.

But the bot would have to be personal and not just personal***iz***ed, i.e., not just fill individual gaps to align the learner to a centralized standardized template of knowledge and skills, but to encourage and guardrail them to pick and choose, follow trails and garden-paths and rabbit-holes, as they desire. (495)

2.6. Cognitive butler?

> *TL;DR I think that AI will mainly serve as industrialized cognition for mass deployment, rather than for personal augmentation and assistance. (472) Would it satisfyingly work also with the small amount of data from a single person's interests? (445) How can emphasis and weighting or bias of what is particularly relevant from either point of view, be communicated to the bot just by verbal prompts and observable behavior? (483) Imagine how great it would be if "he" [the copilot] was so familiar with my unspoken needs that I could even regard him as a my 'butler'! (486)*

I think that AI will mainly serve as industrialized cognition for mass deployment, rather than for personal augmentation and assistance. (This industrialized production does include personal**ized** assistance, such as helping individuals to align with a centrally provided template, e.g. a learning

objectives canon, by identifying gaps and recommending appropriate activities. But I still count that as cheap mass production, and in the worst case it may even be a cheap teacher replacement for those who cannot afford good human teachers.) (472)

Will it satisfyingly work also with the small amount of data from a single person's interests? What about their personal goals for which there is much less data available? (445)

When the target is modelled after a centralized template, there is no mutual influence possible or necessary, and no emphasis and weighting or bias of what is particularly relevant from either point of view is desired.

it is *not* possible to communicate one's personal weightings to the AI via verbal prompts that are limited to explicit ideas and exclude the tacit knowledge. (483)

Imagine how great it would be if "he" [the copilot] was so familiar with my unspoken needs that I could even regard him as a my 'butler'! (who enters the room without knocking, and is allowed to interrupt me with intrusive suggestions that are otherwise blocked.)

I would probably be a difficult client also when my 'butler' tried to read my notes, since I make typos and sloppy cryptic wording and abbreviate so much that my notes almost contain a 'private' language (which I know doesn't exist, according to Wittgenstein). (486)

2.7. Co-learner?

> *TL;DR There is an underlying pattern of Input – Output in the perceptron layers and the End-to-End Machine Learning Workflow, that doesn't seem to leave room for something like reciprocity. Like the famous "unity of research and teaching". Or Howard Rheingold's long-standing practice as a "co-learner" teacher. I think every feedback has the (however small) chance to generate genuine new insights. (458) How can AI get to new insights? After all, AI learns from data of the past and 'ruminates' them. (445)*

There is an underlying pattern of Input – Output in the perceptron layers and the End-to-End Machine Learning Workflow, that doesn't seem to leave room for something like **reciprocity**.

Maybe my idea of Higher Education is too idealistic, based on Humboldt's ideal (see an old CCK08 post) that university teachers and students should learn together, in a community of curiosity and the unity of research and teaching. Or Howard Rheingold's long-standing practice as a **"co-learner"** teacher. I do think that even the unique questions, misunderstandings, or surprising/ outstanding elements for highlighting and annotation, of each new student generation, can influence the teacher towards new insight, even if just by a bit of questioning of old 'matters of course', or just a bit of refocussing.

I think every feedback has the (however small) chance to generate genuine new insights. However, this will probably not pertain to the specialized domain that the robot teacher is trained for, but probably rather come from distant associations, and I cannot yet imagine how AI would implement and handle such extra-domain input. (458)

(Similarly: how can AI get to **new** insights? After all, AI learns from data of the *past* and 'ruminates' them. And as far as I understand it by now, the basis of much of machine learning is the sort of relationships that combine concepts that somehow belong together, within a frame/ script/ scheme, e.g. by 'co-occurrences' of, say, a word on the same page. But can it also come up with types of links like metaphors, or by distant associations? Also, novel data are probably available only in much smaller amounts — which combines both of my above doubts.) (445)

2.8. Robot carer?

> *TL;DR I think there are many cases that only work with a genuine human. E.g., learning by early imitation, shared gazing and trust, relationships of care and coaching and fostering independence. (483) A human coach knows whether the client/ student can bear another challenge, and he signals that he believes that the client might succeed. The client, in turn, needs to trust that the coach really believes this, otherwise the encouragement won't work. But will the client trust a robot coach? I don't think so. And even less so, a robot carer won't gain the trust of a vulnerable 'cared-for'. (455) What the active listening of a 'one-caring' can recognize as the unspoken wishes of a cared-for, is probably not always recognizable by a machine, simply because the cared-for will approach the machine differently. (457)*

I would *not* want to have a robot carer for solace — but why? (458)

Some people may not need or want a personal counterpart to engage with. However, there are many cases that only work with a genuine human. Learning by early imitation, shared gazing and trust, relationships of care and coaching and fostering independence, are all relying on that the other is a genuine human as well. And genuineness is typically recognized from an individual personality as opposed to a templated automated mass instance (unless betrayed, which is why mandatory labelling is the most important part of AI transparency, see below). (483)

E.g., human learning and imitation sometimes depends on genuine human partners: In a study by Andrew N. Meltzoff (whose office kindly sent it to me), titled "*Understanding the Intentions of Others: Re-Enactment of Intended Acts by 18-Month-Old Children*", they experimented with imitation from human vs. inanimate agents and found that "*Children showed a completely different reaction to the mechanical device than to the person*". As I wrote in my predictions for [2021], people are craving for the real, the genuine and the authentic, and this is, for me, the cue that AI's role of personal coaching of unique individuals, will be limited. (450)

Take coaching first (to exclude the additional aspect of vulnerability and dependence of the 'Cared-For' that Nel Noddings so thoroughly described). A human coach knows whether the client/ student can bear another challenge, and he signals that he believes that the client might succeed. The client, in turn, needs to trust that the coach really believes this, otherwise the encouragement won't work.

Now an AI is said to be able to determine whether the student's performance warrants another challenge. But will the client trust the robot coach? I don't think so, unless he is betrayed and gaslighted and told that the robot is a human.

So I think a robot coach cannot help growing self-confidence. And even less so, a robot carer won't gain the trust of a vulnerable cared-for. But for the empathetic feeling to develop in the one-caring, this trust is crucial — it's a mutual development. It's a chicken and egg thing, as Sherida said in today's live session. (455)

Another takeaway of Noddings's work: The one-caring should not act without the expressed wish signalled by the cared-for. So this seems to be once again a matter of pull vs. **push**, which is so important in many tech-related issues. However, in the context of vulnerable and dependent persons, this interplay of request and response, poses yet another subtlety in that the cared-for may be hesitant or embarrassed to explicitly express a need, and so the task of the one-caring is even more difficult, to still recognize the wish by careful active listening, and still *not* overriding the other by preemptive patronization.

(Maybe technical self-service can mitigate some of the embarrassing sentiments, too. When self-service supermarkets arrived to replace the mom-and-pop groceries, one of the success factors was

that customers did not have to be embarrassed when they needed time to decide or if they did not know how to pronounce a product's name. So the reduction of human attendance had at least a tiny welcome flip side.)

However, what the active listening of a one-caring can recognize as the unspoken wishes of a cared-for, is probably not always recognizable by a machine, simply because the cared-for will approach the machine differently. (457)

Also, I think that the careful active listening by the one-caring to the cared-for may indeed occasionally entail that the former learns from the latter, for instance the valuable perspective of a very old person.

This intergenerational mutual learning may be more frequent in environments where fostering independence is important (infants, elderly, HE students, or general critical literacy), still rare but crucial. (458)

3. Ethics

Let them sort, not rank

3.1. Parallels with AI

TL;DR There is this assumption about our own ethics: that it is based on generally valid principles which can be formalized into rules for the AIs. (368) Now the "care perspective", in the immediate situation, is different. (411) One parallel between care and AI is how the 'One-Caring' knows what to do, without rules and without being able to explain it, by recognizing. Another parallel is how the One-Caring learns their ethics: not from central authorities, but rather from individuals in one's close proximity, from examples, via ripple effects, or 'contagion'. (455)

While GOFAI tried isolating propositions, modern AI needs multiple simultaneous data points. And while traditional male philosophers tried to tackle phenomena by rules and from the outside, care respects the circumstances and the inside. "And just so: While good old fashioned ethics (GOFE) tried isolating propositions, modern ethics needs multiple simultaneous data points. And while traditional male philosophers tried to tackle ethics by rules and from the outside, care respects the circumstances and the inside." (Downes)

Often the association pops up that the AIs will be very smart and autonomous, and we hope that they will *want* to behave ethically. And within this idea, tacit assumptions are hidden. One is about our own ethics: that it is based on generally valid principles which can be deduced from reasons, ideally in a scientific way, and can be formalized into rules for the Ais. (368)

Now the "care perspective", in the immediate situation, is different. The decisions and criteria may be learned from others who were in a similar situation and whom we may ask about how they would decide here. (411)

There are **parallels** between care and AI. One parallel is how the 'One-Caring' (as Nel Noddings called them, see Jenny Mackness' wonderful notes) *knows* what to do, without rules and without being able to explain it, by recognizing.

Another parallel is how the One-Caring *learns* their ethics: not via centrally provided templates, principles, etc. or from central authorities, but rather decentrally from examples, via ripple effects, and perhaps like ebb and flow.

Can we **apply** the parallels and differences between AI and human care or learning? One thing is that understanding AI might help understanding how learning works, incl. how to learn ethics. Not via *rules* as in previous generations of AI ('GOFAI', good old-fashioned AI) but via *patterns* and recognition. Thus the old Computer Theory of Mind can be replaced by a more actual metaphor, an AI theory of mind. I find this very plausible (not least because during CCK08 I thought we need an 'internet theory of mind'). It also means a new contribution to the Philosophy of Technology, to see the possibilities and autonomous interaction with AI instead of just the "excessively gloomy picture" (Downes) of its dangers, which indeed fails "the best means for preventing harm."). (455)

In the #ethics21 MOOC, week 8, there was a lot of talking about "society as a whole". In particular, the ethics of the whole society. As 10 years before with the *knowledge* of a whole society, I had my difficulties to get my head around that. So I'll first revisit how it became easier for me to understand it then after Stephen's comment.

I considered approaching the 'knowledge' of a profession or discipline xxx as a newcomer, namely learning how 'they' think and speak and how it may 'feel like' to be one of them. First I might

encounter ‘them’ as some individual new colleague, a ‘you’ in the singular. Then gradually, the commonalities and patterns of their ‘being an xxx professional’, become ever more familiar, and the borders between them begin to blur, and I see them as a ‘you’ in the plural. At the end of this process, the xxxs’ collection ‘as a whole’ contains, strictly speaking, all of them except myself. Then it is only a small step to get from the ‘they’ or ‘you’ to the ‘we’. We all.

Now ethics is similarly learned. From individuals in one’s close proximity. Via ‘ripple’ effects or, as I expressed it in my first vague post, via contagion. Later I learned that this is compatible with connectivism, see ebb and flow. And it has a lot to do with decentralisation, as opposed to central authorities and templates.

Both with knowledge and with ethics, it seems like the ideas ‘spread’ across the interface, or more precisely, *grow* at the interface, between human and human. That’s why it is so dangerous to poison the trust at this interface with fakes. (461)

Learning this ethics takes, so to speak, ‘contagion’ paths, which vary with the decreasing dependence. For infants, the cognitive ‘navel string’ is from mother and parents, later from family and friends, colleagues and communities of practice — the path is the same as for the primordial trust to be seeded and then grown. This percolation path may not yield perfect results and may be slow to change. But it is robust against nonsense from a central, influential source — just as Downes’s “successful networks” promise. (411)

While GOFAI tried isolating propositions, modern AI needs multiple simultaneous data points. And while traditional male philosophers tried to tackle phenomena by rules and from the outside, care respects the circumstances and the inside.

"And just so: While good old fashioned ethics (GOFE) tried isolating propositions, modern ethics needs multiple simultaneous data points. And while traditional male philosophers tried to tackle ethics by rules and from the outside, care respects the circumstances and the inside." (Downes)

3.2. Applying to AI

> *TL;DR “Ethically Aligned Design” ? We software developers cannot effectively influence how technology will be used or abused. But we can alert the appropriate layers about the impending dangers: politics and the voters, and point to critical flaws. IMHO, it is particularly the problem when AI is used to administer, or even create, shortages such as access to the labor market. (368) Perhaps the time has come that an enforceable ‘right of work’ finally enters into the laws. (361) Transparency is a key requirement, and mandatory labelling of artificial communication partners. (368) AI in education should and could mitigate vulnerabilities and oppression, in particular by applications similar to formative (not summative) assessments. (463) I hope that AIs will be useful tools rather than patronizing deciders. If in doubt, as a rule of thumb, they could already do a great benefit if they just do some sorting before the human decision-making. (368)*

“Ethically Aligned Design” is IMHO misleading since we software developers cannot effectively influence how technology will be used or abused. Such an overdone demand would only lead to more frustration and surrender.

However, we can **alert** the appropriate layers about the impending dangers: politics and the voters, and point to critical flaws. IMHO, it is particularly the problem when AI is used to administer, or even create, shortages such as access to the labor market. And the problem of traceability of the algorithms. Transparency is a key requirement. For a start, a minimum transparency should be established by a mandatory labelling requirement of artificial communication partners. After all,

former telephone directory entries here had to carry a symbol if an answering machine was connected. If we acknowledge that trusted individuality is a key human feature, we need to be able to trust our genuine fellow humans.

But transparency and explainability is often difficult. Perhaps GOFAI could now be revitalized and combined with modern AI. Maybe a possible division of labor could also be that some of the riddles of the 'black box' of how AI arrived at its results, could be solved (by explainable AI methods such as counterfactual analysis etc) and then could be fed back into the expert systems. E.g. what are the crucial salient features on a picture of skin cancer that led the AI to a better result than the human expert? (486)

I hope that AIs will be useful tools rather than patronizing deciders. If in doubt, as a rule of thumb, they could already do a great benefit if they just do some **sorting** before the human decision-making. (368)

The most important positive insight for me from the #ethics21 MOOC was that AI in education should and could mitigate vulnerabilities and oppression (see this post), in particular by applications similar to formative (not summative) assessments, and by relieving time pressure.

Also, there was plenty of opportunity to think about the political dimension. The tree vs. mesh structure of society (see this post), the power on the labor market (see this post), the power distribution between end users and those who pay the development (see this post), and whether it will be the poorer students who will be fobbed off with faked teachers (see this post). All of this suggests that we should be very wary. (463)

In particular, regarding the political and to the structures of power and influence.

I would distinguish economic from political power. Economic power (of the few) is mostly derived from the very fact that they *are* few who are controlling some scarce resources (at least this is what my understanding of game theory suggests, which may be outdated since I acquired it when I wrote my graduating thesis in Game Theory in 1977-79). By contrast, political power is the power of the many, who could override and confine the former (at least in a democracy). This political power may be necessary to limit the job market effects — where the owners of AI machines are the few. (444)

I do not believe that the change on the labour market only means that we will have *different* jobs (such that employees just need more education). We must not kid ourselves. The small number of new jobs for programmers and algorithm supervisors, or for managers with the much heralded new kind of soft-skills, cannot outweigh the jobs lost to the cognitive automation, IMHO. Understanding what AI can do now, we will no longer underestimate their competitive power on the labour market.

[A 2018] paper speaks of a *"basis of justified trust and acceptability for users"*, and this needs, IMHO, a palpable guarantee and commitment for jobs. Not job security for existing jobs, but for sufficient jobs. And not just welfare (in a reservation where AI keep us benevolently as well-fed pets?), but sufficient gainful employment.

The paper sets the ambitious goal that Germany shall become a worldwide leading location for AI. If workers have a worldwide leading, justified, trust in their security, this may well happen, because then they will embrace the development rather than procrastinating and resisting it. Perhaps the time has come that an enforceable 'right of work' finally enters into the laws. For example, for every Euro of revenue, x Cents must be paid as wages — with these obligations being tradeable, of course. Otherwise, I think, Luddism will be inevitable, when workers feel like the Silesian weavers when the mechanical loom was introduced. (361)