ARTIFICIAL INTELLIGENCE AND ITS PROBLEMS
May 6, 2023.

Since prehistoric time humans have attributed agency, thinking and intelligence to lifeless things. E.T.A. Hoffman described in 1817 how Nathaniel fell in love with Olimpia, a simple mechanical doll, but the year after Mary Shelley wrote the story of Frankenstein, who made an artificial organism that appeared more intelligent than its creator. It was soon realized that artificially intelligent agents could have drastic consequences on human civilisation: in 1906, E.M. Forster wrote the dystopian Sci-Fi short novel 'The Machine Stops', where all basic needs of humans were satisfied by 'the Machine', and the punishment for trying to cope without the machine was homelessness and death. Humans were busy with various types of hobbies. But when the machine after many generations of humans was about to be elevated to a God it suddenly stopped, with disastrous consequences. A charming feature is the projection of technology of the year 1906 into the far future, like long distance travel using airships. The emerging film industry soon discovered Artificial Intelligence(AI) and an evil AI with human intelligence can be seen in the 1927 UFA silent movie 'Metropolis'. The term AI was coined in 1956 at a workshop in Dartmouth and the same year you could see a humanoid and super-intelligent robot in the movie 'Hidden planet'. The grant application for the Dartmouth session is revealing and the fifth of seven research problems it mentions is the superhot and still open question how an AI can improve its own intelligence. Five years earlier Alan Turing had sketched and discussed the 'Turing test' designed to be used to decide whether or not a machine can think. This is today the second most common way, besides guessing, to decide if an AI is intelligent or can think. Turing's test is unfortunately also based on guessing although in a somewhat constrained way. Literature and movie industry have been ahead of AI, and research has been significantly inspired by Hollywood and Isaac Asimov. But now there seems to be an increasing number of serious applications of AI and media are full of fantastic expectations. Our understanding of AI is now increasing and AI seems, somewhat disappointingly, not very different from other new technologies except that the intensive promotion (mainly using youtube but with significant aid from conventional media) aims to impress the receiver in general terms rather than in promising specific functions.

Deployment of AI leads to numerous mishaps and the applications are less advanced than promised. The intelligence of AI is more like Olimpia's than like Frankenstein's monster's. AI also obeys Merton's law of unintended consequences, and, if the hype of leading experts cannot be brushed aside as embarrassing fantasies, it is time for heavy-handed regulation of AGI development. I am well aware that experts with a liberal mind are against this, one argument being that society is not competent enough to regulate AI research. But once the need for regulation has been convincingly shown, the needed talents will appear. Possibly the current hype is just a mist curtain designed to prevent or delay regulation of narrow AI (i.e., IT development). Such regulation is even more urgent because narrow AI is increasingly used in criminal activities performed by companies and nations besides organized crime, and because of its destabilizing effects on society and its use to inappropriately and often illegally exercise power and to deceive customers. It is used as a higher authority to excuse ignoring basic laws and rules of conduct. Such regulations are under way in Europe and in other places, but they are watered out by strong lobbyists. They have not yet adressed the main problem which is the low security of systems and this reminds us of the early industrial revolution when the new vehicles had no suitable brakes. It is however not entirely obvious that a radically new type of legislation is called for. Spokespersons for the tech industry have a tendency not to know what applicable laws are in place and often suggest that responsibility of an accident (e.g., with a selfdriving car) cannot be handled using existing laws and regulations - such statements

are simply incorrect since cars are already both highly regulated and highly automated, and the popularity of this example only shows that many experts do not know what they are talking about. And the persistent claim of the last 65 years that AI will 'soon' outperform humans in every respect, including stupidity and in appearing human, can by know be abandoned since we will probably anyway notice when this happens. It is more urgent to analyse suitable ways to prevent or delay the clearly visible negative consequences of AI. Certainly, singularity and similar phenomena cannot be ruled out in the distant future, but for the foreseeable future it would be suitable to study things like defining and preventing the use of AI in crime, safe implementations of Asimov's laws of robotics, including law number zero, safe operating systems, and their variants. These tasks are difficult enough, and it is not too late to solve them: regulations are typically implemented after the abuse that motivates them is already a nuisance. After all, Asimov seems not to have been less clever than those who speculate about the singularity now. Rather easy solutions for the urgent safety problems can be found in usability research and in the virtual machine concept developed in the 1970:s such as in the IBM S/360-67, and their deployment can be payed for by shedding the significant load used by todays systems for user tracking, by enforcing suitable constraints.

(note of March 30- April 1, 2023: Finally something happened and a request for pausing AI deployment of large and unpredictable language models has finally appeared on the site of the 'Future of Life Institute' and been signed by a fair number of celebrities. The request is both vague and over-specific, but since the intention is to stop a difficult-to-understand phenomenon this may be inevitable. Also, some of the first signatories of the letter are precisely the persons who I mean, above, ought to be targeted to be prevented from doing certain types of AI applied research and deployment. But the letter may still be part of a secret effort to prevent regulation of 'ordinary' AI. This might have failed when Italy banned the ChatGPT system on the ground, among others, that it spills out sensitive information about persons. This can be expected and can be expected difficult to prevent given the implemented strategy of chatbots to just spit out whatever the statistical prediction methods happen to produce. Clearly, it will happen that a text segment output is a copy of a text segment in the training set, and if it is integrity-violating or copyrighted information it is illegal to reveal it even if it already is accessible on the Internet (or if it was in 2021 when the training set was obtained). It was about time to stop the scam, but it should be noted that the problem here is NOT the super-human abilities of the chatbot but the more mundane problem that the chatbot apparently does not understand that it performs an activity which would be illegal if performed by a responsible human. But since it is not human it is the person/organisation deploying it or using it that is criminal. It is extremely difficult to know when to rely on it and when not. Other problems that led to the ban on ChatGPT are for example that the bot asks the user how old it is and assumes it gets a correct answer, in clear violation of law. My impression is that youtube does the same. Of course, forcing tech industry to follow the law might seem boring and a hindrance for progress as Joakim Wernberg hints at in todays SvD newspaper, but it is a prerequisite for progress. It is dangerous if industry takes its hyped 'breakthroughs' as an excuse for ignoring the law. Those who think that some laws and/or principles on which a law was designed are unsuitable should be invited to work for changes in those laws and principles)

Some of the applications realizing 'narrow AI' are quite interesting and even impressing, so it is clear that their designers are not unintelligent. But the programs themselves are not undisputably intelligent except in a weak metaphorical meaning, since experts are still disputing this. As a matter of fact, intelligence is not among the qualities you infer unless you have been seriously indoctrinated. There are not yet any examples of 'Artificial

General Intelligence' (AGI), but there are many examples of superhuman performance in areas constrained in such a way that the computer has the advantage that we know exactly what it must do: arithmetic, vocabulary and games of leisure with simple rules. All such applications are what is known as 'narrow AI'. But isn´t the recent progress remarkable and indicating a completely new age as many YouTube prophets have proclaimed? No, what is remarkable is only the surprisingly slow progress relative to resources spent shown by AI since serious development started some 70 years ago. The current showcases are not worth the resources that were spent to develop them, but efforts on this scale will always have some surprises in store, like the apparent universal usefulness of simple Markov chain techniques they demonstrate: the answer it produces is the most likely continuation of the prompt under some definitions and constraints, and it is often uncannily reminiscent of an intelligent and useful answer. But this usefulness is an illusion, like the deceptive man-machine interfaces in many computer applications: We cannot yet decide that the answer is useful unless we already know it is. We have despite efforts no broadly recognized way to decide if an AI is sentient, conscious or intelligent. ChatGPT is only a well trained parrot.

Ray Kurzweil, an accomplished inventor and signal processing researcher, has sketched the development of general superhuman AI through 'the singularity', where AI is supposed to surpass human intelligence in all activities since its performance will increase 'exponentially'. This is a bit mystical since an exponential process has no singularities with large magnitudes except at infinity and the exponential increase normally slows down when the approximations describing the process break down. And certainly it is difficult to understand what superhuman intelligence really is, it seems not enlightening to administer a Turing test or a conventional IQ test to the superhuman AI. Indeed, it is rather arbitrary what shall be counted as intelligence of a machine. For this reason it seems futile to pretend that superhuman AI is a fairly well-defined thing and then start speculating when it will be 'achieved'. Kurzweil proposes that the as yet undiscovered solution to the fifth problem of McCarthy's 1955 research grant application can be used by computers to create superhuman AI (The singularity is near, 2005). This is possible but only if a solution is actually found. Kurzweil's story is fantastic but he has worthy adversaries in John Searle, Roger Penrose and Noam Chomsky. Searle (Mind, 2004) has what he calls a naturalistic theory about the interaction between biochemical and mental phenomena which seems to prevent the current generation of AI systems from having real emotions or general intelligence, so more than attention to details and fine tuning will be required before AI is intelligent. He deplores the current misunderstanding that 'it is known' that consciousness and intentions will emerge in a sequence of Turing-computable slightly improving approximations to artificial general intelligence although it is more likely that the current rapid AI development runs down a blind alley - when we reach the end we might not like what we got. Roger Penrose has (The emperors new mind, 1989) ideas assuming that intelligence and consciousness can be explained and possibly artificially realized by quantum-mechanical phenomena not currently fully understood (yes, Penrose claims there are still errors in the current version of quantum mechanics, but Nobel prize winners can afford such luxuries). Chomsky and Berwick (Why only us? 2016) in turn believe that language must be a central part of understanding human-level as opposed to non-human-animal-level intelligence. Some of the brains genes for language (but not the initially promising FOXP2 which we share with Neanderthals) were apparently not evolved for communication but for some other cognitive purpose. There are simple experiments supporting this position, but it remains to pin down the concrete original purpose of the human brain's way to handle language, and a few promising hypotheses are under scrutiny. But this is a controversial idea of Chomsky and Berwick, it is not unopposed although their account is better written than those of the opponents (but not better proof-

read). When the 'Nicaraguan Sign Language' was discovered as the first language designed from scratch in modern times it was hoped that it would lead to consensus around a theory of language development in children, but linguists are first-class quarreling academics so that could not happen. The sign language was created by deaf children whose teachers had failed in learning them lip-reading of spoken Latin-American Spanish. It has remarkably complex grammar of the same kind as existing natural languages have, which would seem to support the UG hypothesis that grammar is severely constrained by the wiring of the brain. It has been a mystery that languages start out being rather complicated and are then becoming simpler as time passes, but here is a possible explanation. Thus, the awkward grammatical structure of languages like German, Finnish and Russian will after eons of use become simpler in grammatical structure (like English and Chinese) because simpler and more useful language constructs than the buit-in ones are discovered which can be learned and give better communication.

There are also analyses by AI experts identifying the most problematic deficiencies in current attempts at realizing AGI by, for example, Melanie Mitchell and by Gary Marcus. In any case, the claims that the singularity is near and the counterclaim that the singularity is impossible both lack support - we are embarrassingly ignorant despite having scattered research and engineering results in large volumes - and the most reasonable position seems to be that the singularity is probably possible but certainly not near. This is an illustration of Orgel´s second rule according to Francis Crick, applied to Kurzweil: 'Biological evolution is more clever than you are'. We do not yet know if engineered AI evolution is more clever than you are. But several top level researchers like the godfather of AI Geoffrey Hinton believes it isn't, one reason being that natural and engineered intelligence seem to be internally fundamentally different and current AI does not show any promise at all in solving several of the hardest problems of intelligence. One or several very necessary innovations are just missing.

The singularity has happened many times in literature and in movies, where the singular AI tends to be evil. On the other hand, societies of believers in the singularity have been studied by anthropologists who found that their members generally regard the singular AI as our savior and good shepherd - and you reach eternal life in paradise by uploading your mind to it. This more positive view seems supported by Kurzweil as well as by some high-tech CEOs. Is this a new religion? Anthropologists (e.g., Beth Singler) see clear signs of this in the focus on superhuman properties like hyperintelligence, defeating death by DNA repair and nanorobots, and ritual meals aiming for longevity to bridge to the coming apocalypse and arrival of ASI. Philosophers (e.g., Roberto Paura) are more inclined to see the untamed utopianism in the phenomenon but also mention the parallell to the predictions of the date for the 'second coming' typical of Christian sects. There is however, as pointed out by the Christian retired math professor John Lennox, a disturbing difference: while Christian doctrine sees God descending to become a human, transhumanists aim to ascend from human to God themselves, like the first Roman emperors. Three implicit criticisms of the singularity religion can be noted: In the Forster novel, the whole story is starkly dystopian because humans have become helpless and unable to resist passivity, and they are orphaned by the singular AI which seems to get tired of them. In the recent best-seller 'Our only life', Martin Hägglund defends the idea of a finite life even with a high degree of automation making full employment by market forces somewhat difficult to attain. He does not explicitly argue against the singularity movement but against the desirability of eternal life, real or only imagined. He gives a deep argument - leaning a bit towards Marxism - for his conviction that money can be successfully replaced by other goals for a finite life (for his Marxism he has been predictably criticized

on the ground that it has not yet been demonstrated possible to administrate conversion to a classless society in a human way).

In media's high tech supported hype there seems to be little expectation for AI to contribute concretely to solve humanity's most pressing man-made problems - climate, poverty, disaster and distress, etc., but the main applications seem on one hand to be games of leisure and replacement of cheap humans by even cheaper computers in the workplace and to plagiarize visual art, music and literature - these applications make service to customers and accompanying promotion much cheaper but maybe only slightly worse in quality (taking advantage of the fact that consumers and employees are normally easily tricked into accepting IT solutions of surprisingly low quality), so it pays off, and on the other hand to take over the control of humanity and give rich humans eternal life while transforming all other matter in the universe to computing devices. AI-experts have claimed that texts produced by chatGPT and similar programs can replace those produced by large numbers of humans - helpdesk crews, copywriters, information officers and management consultants (including the big four companies McKinsey, Deloitte, etc.). It would of course be embarrassing if this turned out true since chatGPT cannot decide if what it reads and writes is correct or significant, it is only an extensively trained parrot. One would expect the replaced professionals to understand their work better - but for management consultants Mariana Mazzucato (The Big Con, 2023) and Douglas Adams (Hitchhiker's guide to the Galaxy) hint that sometimes the difference between a management consultant and a dumbbot is small. Hans-Georg Moeller presents a hoax ('Chat GPT and the paradoxes of our times', Carefree wanderings, 2023) at the expense of ChatGPT when he asked it to review his book on profilicity in a positive tone, and then asked a colleague to ask it to write a critical review. The bot answers by giving two long and detailed texts that shows it like a politician in front of a journalist and video camera, each containing typical positive and negative, respectively, terms it has found in other reviews and giving some general motivations that specifically do not give out any information that can be tied to a specific part of the book. The hoax is humorous but shows a real problem of AI journalism that has not been seriously analyzed yet (and in the meantime Microsoft has banned the system from writing more reviews for public consumption). Since the development teams of the largest search engines seem to work hard to replace hit lists by summaries the quality of search engines will probably deteriorate in the sense of accelerating the trend of not responding with pointers to authentic information the user has asked for, but with a deceptive and uncheckable proposed answer to an unknown question, tailored after what paying customers want the public to believe. This process is already in full swing and it seems odd that the reactions are so mild - the public does not seem to mind. An example (March 10, 2023) can be seen by giving the question 'Is China ruled by law?' to Google and to Bing: Google responds 'Some scholars believe that given China's socialist and non-democratic political system and practice, it is at best regarded as a country of rule by law with law used by the state as an instrument for social control', and Bing answers 'Ja', which is Swedish for 'Yes' so it knows either that I am Swedish or that I am in Sweden. Bing adds that it has found 7 sources but on inspection the sources do not seem to support the answer in any obvious way. Both answers are misleading and the correct answer is no, since the CCP is according to the constitution infallible, in charge, and not ruled by law. But many people claiming that China is ruled by law are blind to international laws of human rights - such laws are typically ignored in many relatively civilized countries,

In Kurzweil's scenario the singularity with a reasonably nice AI can be accomplished by first mapping the connections between neurons of the human brain and producing an artificially as opposed to naturally biologically obtained working copy of the human brain,

predicted to be accomplished in 2025, and then letting it solve the fifth problem of McCarthy's 1955 grant application which sooner or later can lead to superhuman intelligence. This was predicted by Kurzweil to happen in 2045. He is now writing the book 'The singularity is nearer', promised to be published in 2023. We are eagerly waiting to see how Kurzweil will evaluate the quality of his earlier predictions in this book. I would suspect the singularity is not nearer. There is a potential for trouble if, somewhat unexpectedly, humans succeed in creating a superintelligent improvement of an artificial human brain without understanding how it works. Such a brain can like human brains be expected (since it is assumed to work like a human brain) to suffer from mental diseases such as aggressive superhuman schizophrenic paranoia and to start devastating wars - this has happened several times in movies and in literature, but it has also been accomplished by human brains in real life. So the speculations about what 'we' must do in order to prevent disastrous singular AI may be interesting but more important is to realize that larger and more immediate disasters are easy to cause with narrow or even low-quality AI tools and easier than without it. But they are not caused by AI which is only a tool for humans that are not 'we', but 'they', like the founders of Facebook, Twitter, and a number of lesser known subcontracting companies, as well as criminals and leaders of criminal states. Did you recognize the new variant of the NRA slogan: 'guns do not kill people, but people kill people'? It may be considered backwards to make fun of serious AI research, but the research area deserves it: any creation claimed to approach human intelligence should quite obviously tell its user that it is better to aim AI to solve important tasks which humans cannot perform or do not like to perform, rather than tasks which humans perform and like to perform. An AGI must argue against such a position, else it is not an AGI but an AGU. On the whole, the more competent an AI researcher appears, the less concerned he is with the abstract threat of the singularity and the more with mundane unexpected problems of the type showing up in all deployments of new technology.