LLOYD TEVIS
INVESTMENTS, LLC

Securing your future through Precision Investing™

Research Note
A Progress Report on Artificial Intelligence

Introduction

In this research note we make a progress report on AI. AI is a broad term covering activities as diverse as giving machines the ability to recognize and manipulate small mechanical parts, creating assistants which can conduct parts of mathematical analyses, the ability to deliver above human capacities in well defined domains – such as playing chess and acquiring a facility with broad human capabilities such as language and reasoning.

Progress is occurring along a broad front, but two areas in particular have gained much current attention and investment capital. The first is natural language processing and the second is creation of an artificial general intelligence. Accordingly this note will focus extensively on language processing AI.

We first review the foundation technologies. Next we consider applications of the technology. Third we look at the field from an investor's perspective. Finally we close with some personal experiences in this field.

Part 1: Technology

Natural Intelligence

Before diving in to artificial intelligence, we first review some of the things that are known about human intelligence. A key human capacity is a highly developed language facility. Young children acquire language naturally from exposure to language users. They do not require being taught (“programmed.”) Underlying language has considerable structure in the form of grammar and etymologies. For humans language acquisition precedes formal instruction in this structure, but acquiring knowledge of the structure assists with certain aspects of language use. Many other cognitive skills are acquired in a similar manner of initial acquisition refined by subsequent focused instruction – for instance fine motor skills. Not all skills are so acquired however. While humans likely have an innate ability to count and a practical ability to move around in three dimensions, nearly all mathematical ability is taught and not acquired. Similarly, people acquire knowledge that unsupported dishes fall and break. But physics and material science are taught disciplines.

We may usefully compare human intelligence with animal intelligence. Social animals frequently demonstrate communication skills which sometimes approach a level we could describe as languages. Crows exhibit some ability to count. Birds fly without knowledge of aerodynamics and are competent long range navigators at above human levels of functioning. Most of these abilities are, however, either innate (i.e. genetically programmed) or acquired. Transmission of cognitive skills through teaching appear to be fairly rare and very much oriented to practical tasks (e.g. hunting and other means of food acquisition.) Some language skills can be considered culturally transmitted however. Songs of birds and whales exhibit cultural variation. Great apes and dogs have shown ability to acquire human language capacities. In fact, these animals appear to be more adept at acquiring our language than we are at acquiring theirs.

Languages have different ideas of how to structure themselves. English relies heavily on word order with subject verb object (SVO) being the most common structure, as in The man kicked the ball.

Germanic and Latin languages rely primarily on systems of word endings (gender, declension, and conjugation) to express structure and only secondarily on word order. Thus in German the normal word order is SOV (“the man the ball kicks”) but VOS would be perfectly comprehensible (kicks the ball the man), whereas in English it is not. Agglutinative languages such as Turkish string words together to express sentence length ideas in one word, e.g. patting-ones-head-while-standing-one-one-leg-in-a-river Some languages have well developed systems of expression that barely exist in English. For instance Greek has a system of particles that express emotional attitude and which (in written language) might be considered as transcriptions of hand gestures accompanying a spoken utterance. An English equivalent could be considered by different punctuation of the word right (right, right!, right?, right??, right!?) Japanese has a highly developed system for expressing nuances of social respect. In English we have ‘thou’ as an almost forgotten relic of social distinction. Thus the phrase ‘Lord God Thou art mighty” originally signaled a more intimate relationship with the deity than ‘Lord God you are mighty’ would have. Even to a modern ear the ‘you’ sentence still sounds flatter and less significant.

Possession of highly developed grammatical, emotional and social structures makes it difficult for a language to absorb foreign words as they typically do not participate in the natural structures of the language. They also make it more difficult to acquire basic language competence. English, with its fairly weak structures, has a tremendous ability to absorb foreign words. Basic English skills are also fairly easy to acquire. As a result the vocabulary of English is five to ten times larger than other languages. This huge vocabulary gives English a precision, particularly in technical subjects, which other languages have difficulty equaling. However it also means there is a big gap between basic and advanced language competence. More so than in other languages, social and educational class in English is signaled by vocabulary. In German, by contrast, a spoken facility with a verb mood (the subjunctive) would carry such distinctions, e.g. the high born lady says ‘would you be so kind as to point me to the railway station’ whereas the chambermaid says ‘please where’s the station?’

While capable of high precision, English simultaneously possesses high ambiguity as a result of freighting words with multiple meanings. A famous example is ‘the pitcher threw the ball’ which an American speaker of English immediately interprets as referring to baseball and not the possibility than an athlete hosted a fancy dress party. The possibility that a receptacle for milk hosted the party is also not considered, and even when that possibility is noted it comes as a surprise to think of the vessel participating in a sporting event. However in nursery rhymes such would be perfectly possible meanings (‘Hey diddle diddle the cow jumped over the moon and the plate ran off with the Sunday spoon.’) Clearly context plays a vital role in removing ambiguity from ordinary speech.

The ambiguity of English words can be very high. The art historian Kenneth Clark remarked c. 1970 that ‘nature’ had 52 meanings in English. Since then the term ‘natural language’ has developed, so the word continues to acquire meanings. An interesting example of how words acquire meanings is provided by the term ‘panjandrum.’ In the 1770s the English actor Samuel Foote challenged a colleague to a memory test and posed to him a paragraph of perfect nonsense including the sentence “And there were present the Picninnies, and the Joblillies, and the Garyulies, and the Grand Panjandrum himself, with the little round button at the top” About a century later the memory test was applied to a certain farmer, Old McPherson, who was locally famous for his memory capacity. Publication of his story put Panjandrum into wider circulation. By virtue of being repeated by persons who assumed the word had a meaning, the term had by 1920 come to mean ‘a high official of the British government whom no one knows what he does.’ Then in World War 2 a weapon designer adopted the term as the name for a rocket propelled bomb carriage he was attempting to perfect. The weapon proved not fit for purpose and this second meaning would have lapsed into archaic usage only known to history buffs had Google search not lighted upon it and attached it to a photo of the weapon. No doubt pager ranking algorithms will soon note ‘panjandrum’ in the current report and link it to AI. As this example shows, people acquire language through use, but language also acquires meaning from how it is used by humans and now machines as well.

Money is famous for having a triple use as a medium of exchange, a store of value and a unit of account. Language has a triple use as well. It functions as a medium of communication, a store of knowledge and a tool for reasoning. Thus much everyday reasoning is conducted by mentally talking to ourselves. However this reasoning is not especially reliable unless controlled by other means. Our internal conversations are perfectly capable of sounding reasonable but being completely wrong headed. Also conversational reasoning is not the whole of reasoning. The phrases “flash of insight” and “we see that X implies Y” refer to reasoning processes that jump far ahead of conversational reasoning to deeper insights that such reasoning can explain after the fact but generally not arrive at on its own.

Looked at from the perspective of computer technology one notes that language serves as a sort of universal data structure capable of representing any thing, thought or process (‘pitcher’, ‘conservatism’, ‘pas de deux.’) It also is a network in which words stand in relationship to other words along grammatical, etymological, social and semantic dimensions.

Clearly giving computers a facility to work with language would be a real advance. But also clearly it will be a very challenging one.

AI Before Large Language Models

The initial attack on language made by AI researchers took a grammatical approach. One of the first applications attempted was translation from one human language to another. In particular, machine made translation of government and technical documents would have an obvious economic value.

Substantial progress had been made along these lines by 1990. Bilingual electronic dictionaries were developed. Software could analyze sentences and assign words to parts of speech. Complex constructions such as subordinate clauses could be correctly tagged by the more advanced softwares.

But there were obstacles. In the first place, humans refused to use language grammatically. Journalistic and governmental prose in particular, two first use cases, were famous for tangled syntax and generally garbled language. Second one ran up against the problems of ambiguity. A famous example took an English sentence (‘the spirit is willing but the flesh is weak’), translated it to Russian and then translated the Russian back to English with result: “the Vodka is strong but the meat is rotten.” Context would tell us the source sentence is fine for Church and the destination sentence for a restaurant review, but how is one to make a computer understand that? Third computers had even more difficulty than humans in understanding irony, skepticism or sincere belief as expressed through language. The grammatical attack on language processing offered no ideas on how these obstacles could be overcome.

A subcase of the human language translation problem is the problem of translating computer languages.

Computer languages are used by humans to instruct computers how to process data. Computer languages typically have small vocabularies (20-30 words), small completely defined grammars (under 100 rules is normal), unambiguous meaning, one mood (imperative) and a single point of view (prose.) As such they seem stripped of all the difficulties of the human languages. In fact, one could get computers to translate from FORTRAN to C. The translation would be 'correct' in the sense that a computer could compile both texts to machine instructions, execute them and produce identical outputs. But the translations were nearly useless. Humans could read them only with great difficulty. When humans write computer languages they are expressing ideas on how the computer should act. The two languages express ideas differently and those differences mean that the ordinary ways of doing things in the two languages were rather different. The machine translations had no idea of these higher cognitive structures. A C program produced from a machine translation of Fortran would read like English spoken by a foreigner ('I store go now') rather than a native speaker ('I will go to the store now.') This example pointed to a fundamental difficulty in the problem separate from the issues noted with human languages.

At best what machine assisted translation of this era could achieve would be a rough draft which a human translator could clean up. Productivity saving over unassisted translation were too modest for this to become a widely used technique.

Meanwhile a completely different attack on language understanding was being made by historians and military code breakers. Historians possess texts written in lost languages which they hope to read. Code breakers possess texts where the meaning has been deliberately hidden. Both professions discovered that detection of regularities in the text could provide the tool for understanding the texts. Thus

c. 1800 Champillion noted a distinctive way of writing the names of king's in hieroglyphic texts. Steady application of this insight together with external knowledge of the king's names allowed him to assign sounds to hieroglyphic signs. Then possession of a parallel text in Greek and hieroglyphs (the Rosetta stone) allowed him to break the whole language open. In a similar way military code breakers noted that military communications followed set formula (e.g. the header would correspond to a memo header to-from-date structure.) The content of many standard communications also could be inferred (e.g. weather reports, purchase orders, personnel transfers.) Combined with external knowledge (e.g. weather data or freight movements) large portions of the code could be penetrated. Both of these disciplines found that a statistical analysis and an eye for patterns could lead to language understanding.

The First Generation of Large Language Models

About 2018 a break through in natural language processing occurred which is loosely referred to as Large Language Models (LLM) or Generative AI. The basic change was a switch from teaching computers about language with dictionaries and grammars to having computers acquire language through statistical analysis of texts.

The necessary precondition was a large body of texts in electronic form which the computer could learn from. The internet supplied this precondition. Next text would be numerically encoded. For instance one might map a to 1, b to 2 and c to 3. Then a word like 'cat' would become the sequence (3,1,20.) Here each letter is referred to as a token. Long sequences of tokens could be stored in a specialized database known as a vector database. One may define a distance on sequences in various ways, e.g. the distance between sequences X and Y might be defined as $d(X,Y)=|x[1]-y[1]|+|x[2]-y[2]| +\ldots+| X[n]-y[n]|$

Then given a particular sequence X one can ask which sequences Y are close to X. There are several algorithms which are efficient at this task ('the approximate nearest neighbor problem'), among then the HNSW (hierarchical navigable small world) algorithm, the LSH (locally sensitive hash) algorithm and KD tree methods. These algorithms offer distinct trade offs of speed, accuracy and ease of use with evolving data. The basic idea behind this task is that similar sequences owe their similarity to an underlying factor (e.g. perhaps they mean the same thing.) Thus one can hope to ‘extract a meaning’ or ‘recognize a meaningful pattern.’

It turns out this idea sort of works, but one needs more. The second idea is to transform the input. Suppose one is given a long sentence of text. Many of the words will be common and not very meaningful, e.g. ‘the’, ‘a’, ‘is’, ‘it’. Others will be rarer and may convey a lot of the meaning of the passage. For instance just in this paragraph rarer words are ‘transform’, ‘meaningful, and ‘convey.’ Even rarer is the phrase ‘transform input’ and ‘transforming input to convey meaning’ is roughly what we are talking about. So a useful transformation might be to extract these rare words, append them to our original sequence and remeasure nearest neighbors.

All of this sounds clunky and it is. The solution is to adapt the problem to run on specialized chips which can process many operations at once. For instance computing the distance of two sequences of length N in a serial fashion will require 3N operations. But on parallel processing chips it will take 2+log(N) operations. For N=5000 this is the difference between 15,000 and 10 operations. Run time is proportional to operation count so the speed up is about 1500 times.

Large language models pulled together large collections of text, vector databases, search algorithms, transformations and parallel processing chips into a new technology that has proved effective for natural language processing. Some of the things this technology can do are well known

1. It can complete sentences in ways that sound reasonable.

2. Given a question it can search its text collection for answers and either serve them back or show the material it has found.

3. It can combine one and two to create its own answers

4. it can translate from one human language to another with high reliability.

5. It can with some accuracy describe the emotional content of speech.

Casually examined, such behaviors appear reasonable intelligent. Exposed to academic measures of knowledge or reasoning, the LLM may deliver performance between grade school and mid-high school level. This is approximately the age at which human students learn to write essays by building pastiches of information drawn from encyclopedias, which is a reasonable analogy of what the LLMs are doing.

At its heart the LLM is doing linguistic pattern matching. As such it cannot be said to deeply understand its material. This limitation arises when it is asked questions it cannot answer. The LLM is prone to stringing together plausible sounding responses that are completely wrong. When mental patients do this, psychiatrists refer to it as confabulation. Students being tested on material they did not adequately study refer to it as “bull-shitting.” In the computer world the accepted term is “hallucination” as the machine appears persuaded of something that is not there. Similar to mental patients, and unlike students, the computer is not aware that its outputs are unreliable. Accordingly, one of the challenges with LLMs is to detect this situation and get them to respond with “I don’t know” rather than confabulations.

The Second Generation: GLM, RAG and MPC

A second generation of LLMs is coming along. Grounded LLMs (GLM) train their models on texts limited to a certain domain. The model is said to be grounded in that domain. Such models exhibit domain appropriate language use and are more likely to respond in a domain appropriate manner.

RAG (retrieval augmented generation) deals with the problem of private data. LLMs are trained on publically available texts and thus reflect public knowledge. Many applications are interested in private information. RAG is a technique which integrates private information into the resource base of the LLM. Currently this works well with private texts (“unstructured data.”) Progress is being made on incorporating data in databases as well (“structured data”) but the enterprise is still a work in progress. Fundamentally database retrieval is a computational problem rather than a pattern matching problem. LLMs can be given a computational facility but they still lag in this area.

More generally it would be useful if LLMs could integrate to all the existing computer resources. MCP (model context protocol) is a light weight technology standard for making such resources accessible to LLMs. Being lightweight, the technology has a good chance of broad adoption, but it is still too new to assess whether it will prove important to the field or not.

Pattern recognition is a general capability which can extend well beyond text. Pattern recognition can be done with visual data, sound data and procedural data. Similarly the response to a prompt can be images, sounds or lists of procedures just as much as it can be text. These points considerably extend the repertoire of LLMs.

Puzzles

Fueled by a combination of technical excitement and vast sums of venture capital, LLM technology is making rapid strides forward. But some puzzles are becoming evident.

1. Why do computers need millions of pages of training input to acquire language, whereas humans acquire language better with much less exposure? Seemingly humans have built in structures or algorithms tuned to language use which give them much higher efficiency in language processing.
2. Why is it hard for a computer to recognize that it is ignorant whereas humans are highly aware of their ignorance?
3. Humans progress from language acquisition to formal teaching to self education and finally to knowledge discovery ("research.") How can we move computers up this ladder?
4. Can we get a computer to talk to itself? Would it have anything to say?
5. How do humans understand things? It would seem that we build networks of ideas that are parallel to the networks of language, but different.
6. Can LLMs produce a human readable translation of a FORTRAN program?

Clearly there is much to think about, research and try to do. The AI program is well along to becoming one of the great endeavors of our day – comparable to reading the genetic code, arriving at an understanding of quantum physics or mastering flight. Just as those projects are having transformative commercial implications, so too AI is likely to reshape the world in which we live.

Part 2: Applications

Applications: Enhancers, ChatBots, Enhanced Search, and Agents

The easiest application of any new software technology is adding a modest improvement to some other software technology. With natural language processing an example of that would be on the fly spell and grammar checkers. These capabilities check text for correctness as it is being written. They are not entirely accurate, but they do reduce the burden of creating lengthy documents. AI more generally will find hundreds of ways to integrate to existing products and services, making them somewhat better and more effective. Many of these improvements will fly too far below the radar to occasion much notice. People will find they are using AI without realizing it. Currently there are large measures of skepticism and even fear of AI – a natural reaction to the heaping plates of hype that are being served out. As familiarity with "small AI" spreads much of this concern may dissipate.

The first stand alone application of the new technology is the ChatBot. A Chatbot responds to user prompts with natural language – spoken or text. It may hold a degree of context allowing a some what conversational interaction. In commercial applications the principal deployment is in front office (sales and support) operations. The ChatBot can help guide users towards literature or route inquiries to appropriate human staff. Chatbots are already finding broad deployment as the productivity gains are easy for firm's to assess. If the typical client interaction is 6 minutes and the Chatbot, by appropriate screening and routing of calls, can cut the human interaction to four minutes the gain for the firm is a 30% reduction in support cost. Many firms have decided this is a big enough gain for the firm with a small enough burden put on clients that they have deployed Chatbots that are only marginally less frustrating than phone trees. Such deployments are not doing much for the reputation of Chatbots. Unlike phone trees, however, Chatbots are capable of learning and improving from use. We think Chatbots will gradually take over ever more of the support function and customers will ultimately find them to be not just endurable but preferable to all but the best human support.

Internet search has been probably the single most important new technology delivered by the first generation of internet technology. Whereas mail order shopping existed before online shopping, there was no predecessor to online search of note. Any improvement in search is therefore noteworthy. Up to now search has been done by typing a few keywords and the result has been a list of candidate documents which the user had to screen by hand. With enhanced search the user types a sentence and the search function either finds a good match, synthesizes a response or presents possible candidates. As search continues to improve the system will become better at understanding what the user wants – a quick answer to a question, a balanced analysis of a complex issue, a choice of products to meet a need or an annotated list of references. AI will then be able to synthesize an ever closer answer to the users prompt. It is likely prompts will grow longer and more detailed as well. Currently search engine optimization is the most important technique by which firms raise their online presence. As the nature of search evolves, firms are going to have to adapt to maintain and raise their online mind share. Thus AI is going to dramatically impact all firms which recruit prospects online.

Chatbots can chat with you, but they cannot do anything for you. Search engines can do something for you as long as all you want is to be handed a document. Agents take things a step further by, for instance, booking reservations, transferring funds or calling a plumber. In truth all these things can be done by online service providers today. What Agents do is slap a natural language interface on legacy online services. Rather than having the user fill out request forms, the new approach is to engage the prospect in a conversation that elicits the same information. There are several gains from taking this approach. First, the service may be more accessible to casual users. Second, the user may not know exactly what they want or may not know that your firm offers it. A conversational approach allows need discovery and service match to be discovered. Third, complex service offers may be more effectively communicated and offered in this format.

The highest form of an agent is an autonomous agent. The user provides the agent with a very high level direction of what is to be accomplished and perhaps suggestions on how to accomplish the goal. The agent elaborates a plan for achieving the goal and then executes it. The plan may not be fully realized from the start. The agent may need to make adjustments and corrections in the process of execution or solve a series of subproblems. The obvious example of an autonomous agent is a self driving vehicle. The user directs to vehicle to a certain destination and perhaps provides such guidance as whether toll roads are to be used or not. GPS navigation provides an initial route plan which the vehicle drives along. At each point it must adjust its throttle and wheels to integrate with other traffic, obey traffic laws and deal with detours, slick roads and other departures from standard conditions. Today autonomous vehicles exist for both heavy vehicles (freight trucks, tractors, construction equipment) and for passenger cars. Heavy vehicles have attained fully autonomous operation in controlled environments (farms and industrial facilities.) They are under autonomous control on highways with safety drivers aboard. Passenger cars are offering autonomous taxi service in select cities with safety drivers remotely available. We expect autonomous operation to be routine within a few years and to be broadly deployed in about a decade’s time. The gains will be increased mobility for non-drivers, lower transportation cost and some reduction in traffic jams as autonomous vehicles are capable of maintaining good flow in congested conditions better than human drivers can.

There are other examples of autonomous agents. In business management agents can review sales figures and customer feedback, develop market segmentations and propose ways to improve a firm’s offer to better fit demand from the different segments. Firms currently go through that thought process with a fairly slow moving series of human driven analyses, consultations and decision taking. Firms that successfully automate this process will gain competitive advantage versus peers - at least in markets with rapidly evolving demand preferences. Similarly at firms which operate large quantities of capital equipment autonomous agents can review operating logs, maintenance and repair records and service demand forecasts to schedule preventative maintenance to optimize the return on the capital plant. Currently firms likely rely on experienced plant managers for this level of control. The result is disruption in control when key personnel retire, get sick or are promoted. Such normal incidents can be disproportionately costly if the result is that an expensive or vital piece of equipment damages itself due to disregarded maintenance and perhaps shuts the whole plant down. Supplementing human experience with agent oversight can result in returns disproportionate to the cost of the agent.

We are more challenged to find examples of autonomous agents in the consumer sphere, but we clearly see how they might be deployed. An important advance in the treatment of diabetes in recent years has been the development of sensors which permit real time monitoring of blood glucose. This permits patients to see how their body responds to food and adjust both their diet and insulin to maintain adequate blood glucose levels. Better control leads to slower progression of this systemic disease. While blood glucose requires an invasive sensor that only all people will regard as worthwhile, recently sport watches that provide non-invasive monitoring of various vital sign have become popular. It is easy to see how an autonomous agent could monitor this data and make ongoing recommendations for diet and exercise to optimize health maintenance. The natural first adopters are performance athletes and fitness enthusiasts. But it is perhaps not too far fetched to imagine that such practices could become a societal norm if better heath and fitness outcomes actually result.

The same idea of optimizing life processes would apply to finances and education. We will discuss finances later. On the subject of education we note that the current approach is to treat education as a mass produced good catering to the median consumer. Both fast and slow learners are poorly served by this approach. Again intervention of an autonomous agent could potentially optimize the educational process to deliver individualized study programs with better outcomes for students.

Historic Patterns: Productivity, Empowerment, Value Creation and Crossing the Chasm

There is at this point about 50 years experience with commercial development of software. This experience has provided certain insights as to how such technologies are adopted and diffuse. It seems unlikely that AI softwares will majorly depart from these established patterns.

Typically initial adoption requires the software to deliver something of clear value to the first adopter community. Several qualities appear time and again as the item of value: productivity, empowerment and value creation. Productivity gains allow adopters to either process existing workloads more quickly (a cost saving) or to expand the workload that can be handled (potentially a revenue increase.) Substantial productivity gains may allow a restructuring of processes which typically results in large cost savings. Empowerment allows users to escape constraints placed on them by existing ways of doing things. For instance, corporate executives prior to the PC revolution were constrained to rely on corporate typing pools to prepare documents and central accounting to prepare financial analyses. This dependence both hobbled executives and placed deadlines on them that they had to conform to if they were to push work through the production process. A PC with basic word processing and spreadsheet capability allowed them to bypass these corporate choke points. As a result they gained better control over their work and their work calendars. It was this empowerment which led them to champion adoption of PCs by their firms. Finally a straight forward economic case will always be a valid reason for adopting software. Particularly in situations where value creation can be a different order of magnitude from the cost of the software this is an easy case to make. When softwares are first introduced, however, the documentation of net value creation may initially be lacking. It generally takes some experience with the product to understand ts core use cases and the net benefits it can result in. This circumstance typically limits the value creation road to adoption to the situations where the value creation is most compelling.

AI covers a wide range of activities currently at different levels of maturity. For chat bots and capital equipment management the case for compelling value creation appears to be there. By contrast, a health management application would likely rely initially on an empowerment road to adoption and would only be able to articulate a value case after substantial experience had built up. We have previously noted spell and grammar checkers as productivity enhancers which raise the competitive profile of word processing software that incorporate these features.

Typically software is first adopted by a user community which has some special reason for doing so. Often they are poorly served by existing products or they are looking for a source of competitive edge in an ongoing business. A software provider can grow to some extent by serving the needs of this community. But ultimately it typically hopes to grow beyond the first adopters to serve the mainstream community. This transition can prove challenging – so much so that it has sometimes been referred to as "crossing the chasm." Typically the mainstream community has both different motivations and different requirements. Often the mainstream is looking for mature products with well documented but not necessarily dramatic business cases. This is a different mindset from first adopters who accept immature products if they offer evident gains in productivity, empowerment or value. On the requirement side mainstream adopters often pay more attention to sales, support, reliability, security, compliance and brand factors than do first adopters who typically focus on the core product qualities to the exclusion of these peripheral matters.

Currently AI is currently very much in first adopter mode. It is receiving substantial use within the software community itself, but in other industries it is still at earliest experimental/fact gathering stages. One clear obstacle AI must overcome is the problem of hallucinations. Another is challenge it must meet is deploying its capabilities in a valuable fashion for each particular domain. Grounded LLMs and RAG are important tools for creating that domain localization, but they are tools for achieving localization not localization itself.

Commercial Ventures: Infrastructure and Tools

Turning to commercial activities, we find the most mature developments in supplying the ingredients of AI systems.

AI requires parallel processing chips. NVIDIA currently has a dominant position in designing the relevant chips. It initially began developing parallel processing chips for graphic applications. Users with compute hungry tasks realized the applicability of NVIDIA's chips and NVIDIA began supporting this new use as well. AI emerged as one of the important compute hungry tasks. The implication is NVIDIA's chips were not initially designed to support AI. Now that the computational needs of AI are well understood, challengers could potentially challenge NVIDIA by designing chips specifically for AI. Given its huge resource and IP lead, however, NVIDIA will likely succeed in staying ahead of such challenges. Besides designing chips, they must also be manufactured. Here Taiwan Semiconductor holds a dominant position as the independent chip foundry. While not unassailable, Taiwan Semiconductor also is likely to maintain its dominance for some considerable time.

In terms of vector databases, there are a number of providers none of whom seem to have a compelling technical or market edge. Pinecone and Weaviate take the proprietary route, while Milvus, Qdrant and Chroma take the open source route. The standard relational database providers also support vector data to a degree.

In large language models there are also several providers. Developing large language models takes enormous compute resources, so the LLMs are likely to remain a small oligopoly unless a technical breakthrough comes along to shuffle the deck. It seemed that the Chinese company Deep Search might have pulled off such a breakthrough. They achieved a first product by artfully building on LLMs created by others. However, it remains to be seen if they can maintain their edge in a rapidly evolving domain. Probably the GPT models (produced by OpenAI) are the most widely used LLM. However Claude (produced by Anthropic) , Llama (produced by Meta), Gemini (produced by Google) and Grok (produced by X) are all competitive. Which is the current "best" tends to depend very much on the measurement used. There are in addition Mistral (French oriented), Qwen (Chinese oriented) and BLOOM (science oriented.)

The next layer up is AI frameworks. The goal of these frameworks is to provide access to AI tools and methods to programmers coming from the general developer community rather than the AI research community. Primarily such developers are oriented towards creating AI applications and the frameworks aim to facilitate that. The first frameworks were actually developed initially to support general data analysis/machine learning applications. Three of the leading exemplars of this class are TensorFlow, PyTorch and SciKit. PyTorch was initially developed at Meta, while the other two come from Google. All are now open source. Typically they are used from the Python programming language, which itself is the dominant general purpose programming language within this community.

The next generation of frameworks was specifically developed for AI projects. The first exemplar of this class was LangChain, an open source offering from the company of the same name. A number of rivals to LangChain are working to gain traction. Some of these are provided by the LLM suppliers (e.g. LlamaIndex, AgentGPT, Flowise.) Others are independent (e.g. Haystack and Hugging face.)

Beyond frame works are tools oriented towards building specific types of AI applications. For Chatbots we can note Botpress, Botsonic and Intercom. Specifically for AI agents we note CrewAI, AutoGen (from Microsoft), Beam AI and Adept.

This survey is intended only to give a sense of the landscape and not to endorse named vendors. There are many other vendors which have not been specifically mentioned. As one moves higher in the technology stack it gets increasingly difficult to perceive which providers will be long term players.

Commercial Ventures: Applications

In this section we note some current applications of AI. Again the effort is mainly to give a sense of what is available. Some of these might be considered pure AI systems, whereas others are AI add ons to legacy systems. We do not consider that an important distinction. The real question is does it do something useful and well.

Communication Activities

1. Graphic Work

a. Adobe Firefly – AI front end to Adobe's Illustrator and Photoshop tools
b. CanvaAI – AI front end to Canva's tools for creating marketing/presentation visuals
c. DallE3 – Generates images based on verbal descriptions
d. Midjourney – image generator with a different style from DallE3
e. Spline AI – generate 3D models

2. Video Work

a. Filmora – AI enhanced video editor
b. Runway – generates videos with focus on creative effects
c. Synthesia – generates videos with a particular application to instruction videos

3. Sound Work

a. Aiva – music generator
b. ElevenLabs – provides range of distinctive voices
c. Suno – song writing assistant

4. Presentations

a. Beautiful.ai – create slides for presentations
b. Loom – team oriented presentation builder
c. Tome – assistant to create business presentations

5. Writing Assistants

a. Grammarly – grammar checker and editing suggestions
b. superwhisper – dictation taking tool

Education & Research

1. Education
a. Calm – teach meditation skills
b. Duolingo – tutor students in a foreign language
c. Khan Academy – tutor students in academic subjects

2. Enhanced Search/Research Assistants
a. Consensus – research assistant targeted at determining consensus of opinion on a topic
b. Deep Research – researches topics on web and synthesizes

results with citations
c. Harvey – research tool specializing in law, regulation and tax
d. Mem – helps organize research notes
e. Perplexity – web search with summarization and citations

Business Productivity

Automation

a. AI Agent – create an agent
b. Zapier – automate business tasks

Get A Job

a. Lensa – help find a job
b. Teal -resume builder

Marketing

a. Jasper AI – tool suite for product marketers
b. Vista Social – helps manage social media channels

Meeting Tools

a. Avoma – analyzes meeting minutes
b. Fathom – takes meeting minutes and prepares summaries
c. Nyota – meeting scheduling and minutes

Organizers

a. Shortwave – email management
b. Sparkle – organize files on disk by logical contents

Project Management

a. Asnana – AI add on to project management suite

Other

a. xMate – digital friend to chat with

Part 3: Economics

The Investment View: Assets, Moats and Opportunities

In the past 6 months venture capitalists have invested approximately $120 billion into AI. In fact AI has attracted about half of all VC investment. While harder to track, corporate investment also has been heavy. This investment flows both through internal projects and sponsorship of external firms.

Several perceptions are driving this flow. AI is broadly assessed as a foundation technology comparable to the PC and Internet. The perception of wide open opportunities with the chance to establish new franchises exist. Another perception is that firms that "miss this boat" could suffer competitive impairments in their core business. Finally, there is the perception that franchises which looked to be sewed up by existing firms may be open to competitive attack. Established firms feel they should do "learning" projects to see how they can use the new technology. Venture firms hope to be first to "crack the code."

In evaluating possible AI projects investors come back to the same questions that every prospective investment raises:

1. assets – what unique asset or edge does this project bring to market

2. moats – assuming there is something here what defensible moats exist that will let the firm gather the fruits of its labors

3. opportunity – how big is the opportunity here

while the questions are the timeless ones, the answers are specific to AI.

The most obvious assets are hardware assets. As noted NVIDIA and TMSC currently have the best positions. The moats defending those positions are the difficulty of the technology, the patent protections the firms hold, the momentum they have established and the resources they can commit to maintaining their edge. These are not unbreachable moats, but they do ensure that dominance will be sustained for a few years at least and more likely for a few generations of technology. The opportunity is to participate in the global growth of the technology. However only a few percent of the value chain is likely to be realized by the hardware suppliers.

Turning to the LLMs one sees rapidly evolving technology where only a few firms have the capacity to play and drive the industry forward. Know-how in motion is the key asset and momentum is the main moat. Again the opportunity is to participate in the global growth of the technology. In contrast to hardware, the basic methods are well known and patent protection is not effective. As a result the market is split among more hands. Somewhat offsetting this, the firms can offer differentiated services and retain somewhat larger shares of the value chain.

The next layer is tools. If traditional software proves a useful guide here tools can be a stable but smaller business. For it to prosper the tool users must be prospering and so it itself is not a driver of prosperity. The key assets are knowledge of the problem domain and skill in crafting popular tools. The limited potential in the sector dampens competitive fires to a degree and allows tool makers to hold their constituencies.

The next layer is applications. Again if history is a guide applications in aggregate will be where most of the wealth creation occurs. Some applications will have broad applicability and give birth to large firms. Others will be more niche opportunities. The key asset is the intersection of knowledge about the application domain and about the AI technology. Momentum is a key moat. Special know how can also be applicable in certain settings. Category dominant forms typically enjoy excellent margins which allow them to sustain momentum, protect their dominance and also produce free cash flow. Also ran firms generally work hard to keep up. Thus, each individual application category typically looks like a quasi-monopoly or a small oligopoly.

Where AI differs most from earlier generations of software is in how one sizes the opportunity. The previous generation of software was SAS ("software as a service") – server based software providing online services. Most SAS business software sells itself on a productivity gain basis. These gains were at the 20% level for the most part. Once the gain was proven the sale was fairly easy but the value to be captured was limited. AI offers potentially larger productivity gains. Today a business function might be staffed with a full time manager and 5 staff. The cost would depend on the industry, but the fully weighted cost might approach $1m per year. Introduction of AI could potentially result in this work unit being replaced by two technical specialists assisted by AI. Fully weighted cost might be $350,000 per year for a productivity gain of 65%. Per employee salaries will actually have risen, but headcount will have decreased and that decrease will result in secondary savings in facilities, benefits, and general support and management costs. These secondary savings can be quite meaningful in aggregate across a firm. Thus in sizing the AI opportunity understanding its capacity to drive process re-engineering is critical. In particular, with autonomous agents there can be the capacity to change the basic economics of a business – costs that previously scaled with the growth of the business may be converted to fixed or at least slow scaling costs. Fundamental change of this sort can change the viable offers in a marketplace and displace incumbent ways of doing business. In that case there may be the opportunity to capture a significant share of industry revenue.

Reviewing the applications listed above we note the substantial presence of AI in communications and marketing. Indeed marketing seems to be one of the business functions where AI is having an early re-engineering level impact. In fact, marketing is an activity driven by data analysis but high reliability is not a requirement of the field. These qualities make it a good fit for the current generation of AI. An AI which occasionally hallucinates may be no worse than a marketing executive whose brain waves are sometimes fueled by one too many martinis. As an area of early impact, tracking the impact of AI on marketing may provide early insight into how rapidly AI is changing the business environment.

We believe, however, that AI will have broad applicability and be transformative in many domains. Education, document research, plant optimization, health monitoring, and financial services are all areas where AI can have major impact.

The Hubbub: Hopes, Fears and Challenges

AI has produced a great deal of excitement which feeds many hopes and fears. Technologists speculate that the point at which machine intelligence surpasses human is near at hand. This might on the one hand pave the way to faster scientific and technical advance. On the other hand it might empower abusive and stupid human activities. The advance of AI is creating trepidation in the broader population and particularly among office workers. They wonder if their roles will be assumed by AI and whether process re-engineering will terminate their careers. There is also the concern that entry level jobs are being reduced in numbers with the consequence that middle class careers will be out of reach of their children. Some investors speak confidently of the AI sector of the economy being larger than Germany and perhaps as large as China within five years.

Both the hopes and fears rest to a degree on the perception that technological improvement will be large and quickly arrived at and that commercial adoption will be broad and rapid. The natural question to ask is what data do these perceptions rest on and how informative is it.

There are certain scaling laws which show measured improvement of AI versus various resource inputs (e.g. core counts, size of training corpus, training time etc.) These laws have held steady over a wide range of resources sizes. At least for near term extrapolation they give a sense of what performance might be expected from the next generation machine. Mostly these results appear in the setting of unsupervised learning where the machine is exposed to a large quantity of input and is left to find such patterns as it may. A different scaling law appears to apply to supervised learning. In supervised learning the operator attempts to teach the machine a pattern by providing curated data and possibly providing feedback on the machine's performance. In this setting a sigmoid scaling appears to apply. Initially there is little response to training. But once a threshold is passed improvement is proportional to effort up to a point. Then saturation appears to set in and further increase in effort produces only minimal further improvement. The presence of saturation effects warns us to be cautious when extrapolating performance forward by multiple machine generations.

A glance backward at the history of AI also is a source of caution. There was almost twenty years of steady work on the grammar approach to language processing before it became clear that new ideas were needed. It took a further 10 years for those ideas to emerge and another 5-7 years to test them out sufficiently to see that they constituted a real breakthrough. In terms of such goals as achieving robust AI reasoning, attaining broad self education or creating an artificial general intelligence it is quite likely that additional novel breakthroughs will be required and so it is impossible to estimate with any degree of confidence when those objectives might be met.

We think a more meaningful forecast might be given as to what sort of AI can be constructed over the next 5 years. We proposed a year ago that by 2029 it should be possible to construct an autonomous agent which could conduct a chemistry research program with a set fairly narrow purpose, some meaningful cognitive depth and over a very wide range of cases. This would be an example not of a artificial general intelligence, but rather of a specific intelligence purpose built through integration of a number of component technologies.

We will update our forecast for 2030 to state will not only be possible to create such devices by then, but that a number of different autonomous agents of that nature will have been created by that date. We think such devices may locally impact careers and job opportunities in certain domains. For instance entry level chemical research may have shifted from swirling chemicals in beakers to programming and validating the work of robotic research assistants. However massive economy wide impacts seem unlikely to occur in that time frame.

One conceptual shift we do see occurring with widespread use of autonomous agents is concepts of how such agents should be supervised. The natural first idea is that of the human safety driver who watches the agent perform and is prepared to step in to correct poor performance. We find this concept natural but naive. One problem with it is that it requires humans to maintain high vigilance and be prepared for bold action through long tedious hours of routine operations. Humans are not very good at such tasks. Second it assumes one can turn the agent off and revert to old style operation. But autonomous agents will quickly be running at performance levels which cannot be replicated by “manual operation.” Switching from “automatic” to “manual” will results in unacceptable performance costs. Instead one needs some way for automatic operation to be continued in a ‘safe mode.’ Third, the idea of a ‘safety driver’ is oriented towards catastrophic failure of the agent. The more common cost of a poorly performing agent will be extended periods of running at a performance level close enough to optimal that its deviation from optimal is to not readily observed but far enough from optimal that substantial costs are accumulated. Complex factory processes are subjected to statistical quality control regimes to handle this sort of problem. Similar methods will likely be needed for autonomous agents.

As experience with autonomous agents grows, we expect the limitations of the 'safety driver' approach to become better understood.

Part 4: A Case Study

One of the first fields to apply machine learning and AI techniques was investment management. In this context the activity is known as "quantitative investing." The present author has been intimately involved with this activity for 35 years and has had the opportunity to see several waves of technology travel the adoption curve. Recounting that history provides a case study that may prove informative to how adoption will proceed in contexts which are just beginning to apply these methods.

Academic theories as to how investment management might be made into a quantitative discipline began in the 1940s. It took until the mid 1970s, however, for practical application to begin. The key enabling idea was the development of the risk model. Previously investment risk had been assessed subjectively and only qualitative distinctions were made. The risk model made such assessments objective and quantitatively fine grained. Risk models came out of statistical analyses which for the day were impressively complex. The data sets involved were a few megabytes which meant that they could only be stored on magnetic tape. Estimating a model required many passes of sequential processing in which human operators intervened to mount new tapes at the start of each pass. A total estimation could take months of such processing work. The fitted models would then be provided to users via VAX based time sharing services. The user community was limited to a few teams scattered through major investment institutions. Corporate policies on data security, insider trading and similar concerns generally slowed technology adoption. If this sounds not unlike the world of LLMs today that is no accident. Where the cutting edge is moves but what it is like changes only slowly.

By the mid 1980s awareness of quantitative investing was growing and personnel just joining the industry were beginning to specialize in it. The personal computer entered the corporate world in the early 1980s and by the mid-1980s it had gained sufficient power that it could begin to run risk model based investment analyses. This development made quantitative tools accessible with less involvement by corporate supervisory departments and adoption began to pick up.

Initially quantitative investment management had focused on the portfolio management setting. Here decisions were taken once a month at most and more typically once a quarter. Practitioners were used to hand picking investments and setting their weights in the portfolio. Indeed they saw that activity as a central part of their well compensated job. An algorithm (“the optimizer”) was capable of taking on this task and actually performed it better than the humans did. However the industry only adopted it slowly and over considerable resistance. To get it adopted the technologists needed to create manual tools which would allow the user community to play the “safety driver” role. A few years of exposure to checking that the optimizer had done its job accurately both educated users that they could rely on this tool and that they would have to justify their employment some other way.

By 1990 computer technology had advanced to the point that quantitative methods could be applied to the trading problem. Here the datasets were gigabyte sized. They needed to be stored on optical platters and a mechanical arm would pull the required platter out of its storage slot and place it in the reader when needed. The device was, therefore, known as an optical jukebox. Unfortunately its control software was liable to getting confused and losing track of platters. From time to time it would halt job runs and have to stop to rebuild its platter inventory. This made the device almost as frustrating to use as the manual tape mounts of former days.

Processing all that data was a job spread across several dozen Sun Sparc stations. During the day time the analysts working with monthly data would be using these boxes to conduct interactive statistical analyses of their small data sets. At night the trading analysts would utilize the boxes to run big batch jobs. There were no commercial tools for doing distributed parallel processing jobs, so the trading team needed to create their home grown versions of such tools. Today, of course, "big data" work is supported by numerous commercial tools.

The faster pace of the trading environment also turned attention to analyzing text based news feeds. Grammar analysis of headlines and keyword counts on text bodies were the cutting edge of technology then. Neural nets were also getting some attention as universal pattern recognizers. By 2000 they would be applied to analyzing trading patterns. But I doubt anyone could have foreseen their convergence with text processing at that point in time.

By the early 2000s computers were ready to take on the market making function. At that time market making was done by humans either shouting at one another on wooden trading floors or manning telephone banks in front of computer screens. Traders were known (mostly to themselves) as "masters of the universe" and they were quite confident no computer could do their job. Traders would scribble their trades on paper tickets which the back office staff would cross check with the back office staff at the counter party firms. Illegibilities and discrepancies would have to be worked out in a friendly way after hours. I remember bringing online one of the first market making boxes at such a traditional trading firm. The head of operations asked many how many trades I would be doing in a day. He blanched at the answer and said "that is as many trades as the rest of the firm does." Gesturing at a huge room full of back office staff , he said "I am going to have to double the team and I wont be through with break processing until midnight." I assured him he would not need to add any additional staff and there would be no breaks. Initially he dismissed me as an ignorant boffin, but after a month of running he sought me out to say "you are right there are no breaks, we square your book within ten minutes of market close. I love your business." Within a decade the human traders had been rechristened as "screamers" and they were a nearly extinct profession. Computers had taken over the market making function in its entirety. In retrospect algorithmic trading is probably the first application of autonomous agents to control of substantial economic resources (billions of dollars.) The computerization of trading has also massively expanded the data flows. Today's trading datasets reach the petabyte level. The combination of requiring massive computer resources and highly specialized teams has limited top level algorithmic trading to a small oligopoly of firms. Even here, however, competition gradually trims margins.

The current frontier of quantitative investing is in personal finance. Institutional portfolio management deals with a few portfolios each of which may contain hundreds of positions constituting in aggregate hundreds of millions to billions of dollars. Trading deals with hundreds of assets about which thousands of split second decisions must be taken aggregating to tens of billions in flow . Personal finance deals with thousands of fairly simple portfolios each of which holds just tens of thousands to a few million in dollars but which again aggregate up into billions of dollars. The challenge in this situation is dealing with a high level of complexity as each fund owner has their own set of objectives to fund with the portfolio and those objectives are as distinctive as fingerprints. Computers have a hard time dealing with complexity and that is why they have gotten to personal finance only after handling the portfolio management and trading problems. That is pretty counter intuitive as most people would think the small account was the easy problem and the big account was the hard problem. But history shows that what is hard about a problem may not at first be obvious.

The currently deployed technology in personal finance was designed in the early 1990s. It puts portfolio decision making in the hands of small back office teams, while an army of front office advisers explain the product to the clients. About 2004 initial attempts to replace the front office advisers with "robo-advisers" began. A robo-adviser might make you think AI is involved, but it is not - a robo-adviser is just a brochure server which matches clients up with product literature. Its not an especially impressive technology and that has limited its adoption to the low end of the market. The attempt to deploy actual investment intelligence began around 2006. By 2010 the technology had advanced well enough that it could reliably generate better decisions than human advisers. The banking crisis of 2008 had a chilling effect on technology innovation in this field. That freeze lasted a bit more than a decade. But today autonomous agents for personal finance are being deployed.

LLMs have some interesting value to add in this area. Financial plans have the same problem as diet and exercise plans. They are great on paper but users have difficulty coordinating the micro-decisions of daily life with the "the plan." As a result the plan may stay great on paper but not actually get implemented in life. LLMs make it possible for users faced with such a micro-decision to pick up their phone and ask their autonomous agent for assistance. Making the task this simple and easy is the value LLMs can bring to the table. In the process they importantly improve the probability of good outcomes for the client.

Overall I think this case history is cause for optimism. Technologies that deliver real value get adopted and applied as they should be. Sometimes internal resistance or external circumstances may slow the process down, but they do not change the ultimate outcome. At the same time, things happen at a measured enough pace that individuals have the opportunity to evolve their careers and adapt to the changing world. As compared to the disruption wrought by recessions, financial crises, pandemics or political chaos, technological change is a comparatively gentle force.