The Inverted Matrix: A Vision of Hope and High Aspiration

Thomas Dean[1,2]

Abstract

How might humans, cyborgs and artificial intelligence work together to create a shared vision for the future and what could we do now to make this possible. Some experts believe it is inevitable that humans will one day share this planet with biologically enhanced humans, cyborgs and artificially intelligent robots. We've been told the age of AI might lead to massive unemployment, advanced weapon systems, pervasive surveillance and law enforcement, and even the rise of artificially intelligent overlords. These bleak visions of the future may be possible but they need not be inevitable. The technology of artificial intelligence could be applied to realize a future in which intelligent systems of many sorts create a global framework for planet-wide collective decision making, share in the governance and stewardship of our planet, amicably accommodate their differences and resolve disputes, and ultimately forge a path in which technology and shared purpose replace natural selection as the engines of human evolution.

We discuss some of most relevant advances in cognitive and systems neuroscience, AI and machine learning research leading to advances in technology that could transform society and significantly improve the lot of human beings by unshackling us from instincts evolved millions of years ago to enable us to survive a kill-or-be-killed jungle but are now counterproductive to our happiness and wellbeing. Soon we will have technology that will enable physically unaltered humans to substantially extend their cognitive capabilities, and eventually physically augment their bodies to achieve even more efficient interfaces with machines and expand their capabilities still further. Looking beyond the next decade, unaltered humans, augmented (cybernetically enhanced) humans and nonbiological AI systems will populate earth and it is (theoretically) within our capacity to ensure that such heterogeneous populations co-habit this planet amicably, learn from one another and govern themselves so as to allow all to prosper and realize their potential. This optimistic future is not certain, but will have a greater chance of success if companies begin planning and producing products and services that create an environment in which relevant innovation will thrive, or, better yet, create consortia to independently pursue and freely disseminate the relevant enabling technology.

The truth is that without Human AI as partners, proxies and prosthetic thinking appliances, we — members of *Homo Sapiens* 1.0 — cannot scale our innate social-bonding and decision-making capacity to solve the problems facing us today. Perhaps that seems overly pessimistic, but human history [3, 2], psychology [14] and neurobiology [10] provide little evidence to discourage such a dour outlook. Already, it seems that we are ratcheting up our xenophobic and (parochially speaking) misanthropic distrust of strangers to vilify robots and AI systems. Shackled AI systems are like stockpiled nuclear weapons just waiting to be detonated, stolen, or used as threats to further the ends of dictators, despots and terrorists, and once we can deploy other AI systems to pick the locks and hack the boot ROMs we might as well just hand over the keys. We are warmongering the citizenry, stirring up the troops and rattling the sabers even as the enemy is still waddling around in diapers. It is time that we are honest with ourselves, admit our weaknesses, and, as we have done so many times before, invent technology to overcome our shortcomings and control our destiny.

Affiliations: [1]Brown University, [2]Stanford University

Contents

1 Inverting the Matrix

Imagine a world in which education is universal from grade school to college, on through graduate school and beyond for anyone willing to put in the effort — the cost is free, the opportunities limitless. Imagine a world in which it is effortless to avoid the imperfections in thinking that permeate primate brains shaped as they are by natural selection and the need to survive in a world hostile and full of dire threats. Imagine a world in which everyone can contribute and participate in its governance — a world in which AI systems help to coordinate our activities and consolidate our opinions and aspirations so that billions of people can efficiently arrive at agreements on how to deal with our political, national, ecological and climatological problems. Imagine all of this happening in our lifetimes, culturally, politically and sociologically on a global scale, and, on a local scale, cognizant of the consequences of our actions on our lives and those of our children with a scope that encompasses the here and now and yet embraces the goal of ensuring a sustainable future for an indefinite span of human development.

Imagine a future in which truly personal AI systems act on our behalf helping us to negotiate the often bewildering complexity of the world in which we inhabit by dealing with the constant threats of identity theft, the loss of financial or health records, figuring out what to do when we or our loved ones have to make difficult personal health decisions, filtering the information we receive by sifting out items of interest or utility and fact checking information from unknown or unreliable sources. Imagine how we might contribute to decisions that impact our lives both locally and globally in such a way that our voice makes a difference and the opinions of individuals or corporations with great wealth are treated on the basis of their wisdom and not their vested interests, where unchecked emotions and divisive bias are not tolerated and don't have a chance to escalate, cloud our thinking and destroy our chances of coming to agreement on issues critical to our survival and happiness.

Imagine a world in which learning is tailored to our individual needs and interests at a pace and in a manner that accommodates our particular cognitive strengths and weaknesses. A world in which learning never begins or ends and is offered free and of the highest quality to everyone. Imagine a future in which AI systems are engineered to understand human frailty both physical and cognitive, so as to assist us in health and sickness, when tired or refreshed, when stressed or tranquil, energetic or lethargic, cognitively challenged or intellectually supercharged. Imagine a world with transparent boundaries, in which villages, small towns, large cities, entire nations and continents are able to resolve their differences and make choices that respect and serve the varied constituencies inhabiting each of these geographic and geopolitical divisions and ultimately lead to a level of trust and understanding at which these boundaries completely disappear. Imagine a network of AI systems organized dynamically and hierarchically to serve these constituencies by balancing, coordinating and integrating the efforts of diverse interests, making it their business to understand the particular needs and aspirations of their constituents and represent them within a global network of systems all working to improve the present and create a future that serves our collective vision of a world in which we want to live and bequeath to our children.

Some upon hearing such a broadly sweeping vision might dismiss it as the wishful thinking of technocrats believing that all problems can be solved by technology, but our view is that this is the vision of a future in which technology plays an important enabling role but human beings ultimately make the decisions and direct their future. Some might imagine a dystopian future in which our every move is tracked by intrusive surveillance controlled by an all-seeing all-knowing AI system like Skynet, or our minds controlled by implants and imprisoned in a simulated reality like the Matrix. Here we seek to build systems that attempt to avoid the need for pervasive surveillance by ensuring that all citizens have unlimited access to good nutrition, healthcare, education and opportunities to express themselves and participate in meaningful ways, and to create laws and the means of enforcing those laws that respect privacy while at the same time understanding the trade-offs and recognizing that the cost of limiting privacy and curtailing more aggressive policing will

inevitably require some sacrifices in terms of security and peace of mind.

Our personal AI systems will combine the skills of a teacher, confidante, amanuensis, personal negotiator and life coach. As a cognitive prosthesis, they will augment our innate abilities so that many routine cognitive tasks like understanding inscrutable legal documents, filling out insurance, medical and financial forms and keeping track of names, appointments and prescriptions will become effortless, while cognitive tasks like wrestling with the trade-offs involved in taking steps to ameliorate the global consequences of climate change will continue to be intellectually challenging since they have no easy solutions even accounting for AI assistance in making sense of the science and a collective coming-to-agreement in separating fact from fiction. Decision-making will revolve around what we value and what sacrifices we are willing to make to accommodate the uncomfortable consequences we have imposed on ourselves by failing to come to terms with difficult problems due, in large part, to political, national and economic factions that have applied inordinate and morally unacceptable pressure to profit themselves.

Human beings are rational only with self-imposed control and deep insight into the patterns of thought and proclivity of instinctive behavior that dominate our thought and intercourse with others. Science is unraveling these complex patterns and instincts, but even the psychologists who study human decision-making routinely fall prey to the evolved cognitive expedients that served us well over the majority of our evolutionarily-recent past but are a liability in the world we now inhabit. Personal AI systems will serve as coaches to overcome such deficits and over time we will learn to control these ancient and out-of-place instincts without explicit coaching, the outcome being that a good deal of human intercourse will become more civilized, comfortable and productive — resulting in a world with less social unrest and fewer interpersonal altercations. Having a powerful AI system as a personal coach is not a panacea for global peace or emotional tranquility, but it's a movement in the right direction as humanity takes control of its evolution to drive its destiny.

We are already digitally superhuman — Apple, Google, Wikipedia, etc have seen to that. Kernel, Neurolink along with dozens of academic labs want to improve the interface between humans and machines. We all want to be faster, richer, smarter, etc, while what we need is to become less selfish, parochial and short sighted. It is not enough to know how to solve the huge problems facing humanity. One needs to negotiate global solutions that accommodate and anticipate the needs of billions of people living now and yet to be born. One needs to bring together stakeholders from all over the world, agree on a solution, avoid easy, short-sighted political compromises, proceed together one step at a time, dealing with setbacks, revising plans, matching expectations, soothing tensions, accommodating new ideas and, above all, realizing and compensating for our own shortcomings by using technology to make us better humans.

2 Handling the Truth

You're not going to like this. No one really wants to hear that we are intellectually challenged and emotionally imbalanced. As Jack Nicholson's character, Colonel Nathan Jessep, in the film *A few Good Men* said to Lieutenant (junior grade) Kaffee, "You can't handle the truth!" Well, perhaps *you* can, but it will be hard to swallow that you are one of the main problems standing in the way of political and social progress. You have an antediluvian brain operating in the 21st century. It isn't optimally suited to deal with global problems and can't even be guaranteed to behave well at your child's next PTA meeting.

Robert Trivers in *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life* [14] made the case that one reason we can successfully negotiate in social situations is because we deceive others by masking our true intentions even from ourselves. Without self deception we would inadvertently advertise our true intentions through revealing facial movements and body language. This can be upsetting to hear, especially when you learn that humans engage frequently in self deception, are often unaware of doing so and that it works quite well for us personally as one might expect from a naturally selected trait [11, 1].

The problem is that what makes us so effective in pairwise negotiations doesn't necessarily scale to making decisions on a global scale[1].

There is a large body of knowledge describing the many ways in which human decision making and self assessment is biased and deviates substantially from more rational or optimal procedures. Cognitive behavioral theory [13] and mindfulness meditation [7] are but two of many approaches to dealing with the sort of negative distorted thoughts common among people in general and anxious and depressed people in particular. We are quick to take offense, misinterpret harmless comments as slights, and once angered or anxious our self control declines markedly and our cognitive facilities for judgment and reasoning are diminished [8]. Ostensibly these reactions were important to our survival but today they are deficits.

David Kahneman and Amos Tversky's research was largely responsible for the development of *behavioral economics* which is used to describe how people operate in real markets and deviate substantially from the predictions of expected utility theory. Their description of the *framing effect* explains situations in which people react to a particular choice in different ways depending on how it is presented [4], and, in developing *prospect theory* they describe heuristics for judging frequency and probability that explain how people reason about chance leading to poor decisions and outcomes in many professional circumstances [5].

Kahneman and Tversky [6] articulated the principle of *loss aversion* asserting that the subjective weight of penalties is larger than that of potential rewards. Hence, for example, people should avoid lotteries that offer a 50-50 chance for equal-sized gains and losses since the negative consequences outweigh the positive ones. They proposed that the reason for this bias is rooted in an affective context: "the aggravation that one experiences in losing a sum of money appears to be greater than the pleasure associated with gaining the same amount".

These distortions, deceptions, misunderstandings and heuristics served us well when we were routinely exposed to life-or-death situations in which it was more important to be occasionally wrong and live to see another day than right and some predator's next meal. Now these behavioral, emotional and psychological aberrations all contribute to our flawed thinking and emotionally fraught deliberations, decisions and policies. They prevent us from arriving at good outcomes, lead us into war, create civil strife and tear families apart and sunder nations into ideological factions incapable of civil discourse.

3 Journey Starts Here

It may seem demeaning to suggest that we — *Homo sapiens* 1.0 — probably can't overcome some of our cognitive weaknesses without advanced AI technology, but neither can humans without the aid of construction machinery perform the physical activities required to build skyscrapers or engineers without computers manage the intellectual feats required to perform the calculations necessary to design bridges that span kilometers and carry loads consisting of millions of tons of freight.

If anything, the challenges are even more daunting given our deeply ingrained social instincts optimized for small clannish communities living in the primal savannas of Africa. You can't just put on a pair of rose-colored glasses and fake a smile and expect the members of a neighboring clan to lay down their spears and equitably share the available resources even as they wax and wane according to weather patterns you can't anticipate, quantify or control.

It is not as simple as following some protocol a game theorist has declared to be optimal. The problem is that we can't control our emotions, and once in the thrall of our emotions we simply cannot think clearly. In particular we can't judge how our emotions will cloud our expectations regarding the thoughts and possible

[1] Self deception may be psychologically comforting or even necessary in some cases. It would be surprising if there aren't some decision-making strategies involving deceit that are unavoidably part of complex solutions. Judges, jurors and legislators often weigh the benefit for society against the cost to the individual or special interest group. We might want to deny our complicity in a decision that seems horrible and inhumane for a few even it avoids enormous suffering for many.

actions of another, how they ratchet up our inclinations to strike out causing our erstwhile opponent to take on a defensive posture or escalate offensively.

Having altered the tenor of a conversation by making a derisive comment, we don't know how to gracefully repair the damage and return to a level of civility necessary to continue a constructive dialogue. Adding even one other person to a conversation can complicate matters making it even more difficult to carry out the sort of complex negotiations required to forge agreements across social and political boundaries.

In many cultures such politeness is neither valued or practiced. Indeed it is deemed a sign of weakness suggesting that demanding and bullying may be effective strategies. Such careful control of one's emotions, accurate reading of the other, and willingness to retract and apologize for aggressive posturing are rare even among diplomats and at the same time essential in forging solutions based on trust and mutual respect rather than a careful appraisal of force and the willingness to apply that force.

These skills are rare in individuals and so it is ludicrous to believe that they will emerge spontaneously from an angry crowd of soccer fans or a frightened collection of representatives in the UN General assembly arguing over the proliferation of nuclear weapons, the growing threat of global warming global warming for coastal communities, or increases in sectarian violence and terrorism spreading into yet new venues.

Part of the trouble stems from misunderstanding intentions, assuming the worst, expecting lies, promulgating misinformation, applying zero-sum thinking, and forgetting that, almost universally across human intercourse, losses compel more than gains by a large margin. In retrospect, it is incredible that we have achieved some relaxation of international tensions, a tenuous detente agreed upon primarily to achieve a continuation of commerce upon which all nations depend to provide for their citizenry and fund efforts to modernize and increase the strength of their military might.

So what sort of solution are we proposing here? The solution is predicated on the assumption that we need help in controlling our emotions in guiding our decisions in order to direct our lives individually, and perhaps more critically at this point in our evolution to assist us in interacting with one another so as to evaluate, select and execute procedures and policies for solving the world's most pressing problems.

While it may be that, if successful, this proposal will bring about a new way of governing the world in which we live, that is not its immediate goal nor is it a realistic outcome we can hope for anytime soon. Dismantling the bureaucratic machinery of governments and corporations will only happen if those with vested interests in maintaining such machinery deeply appreciate the advantages of the sort of symbiotic relationship with machines that we advocate here.

What we hope for and believe possible is that enlightened bureaucracies, governments and corporations will provide scaffolding enabling us to build and maintain the infrastructure required to facilitate the complicated interplay of minds, including unaltered humans, cyborgs and AI systems, necessary to create the sort of intellectually informed and cognitively prepared citizenry we believe this world needs to survive into the next century.

4 The Matrix Revisited

In some respects, the world we live in now is becoming more like the movie Matrix. Increasingly, our purpose in life — our ability to contribute to the world in which we live by helping to determine how it's governed and having a voice in defining its values and aspirations is being diminished as the availability of meaningful employment diminishes, corporate and special interest influence on elections rises, and the healthy inclination toward innovation and entrepreneurship is replaced with the desire to succeed at any cost.

The jobs that most people can get are often demeaning or involve little cognitive effort or creativity and soon even those jobs will disappear, and so, like the Matrix we live our lives imagining ourselves to be the actors in the movies, sitcoms made-for-TV serial dramas, and even these often depict worlds such the most fertile imaginations have trouble seeing as anything but brutish and toxic.

So, like those poor unfortunates imprisoned in the Matrix and manipulated by their AI overlords, we live in a virtual world designed to tell us what to think, but with the added injury that unlike most of those living in the Matrix we know that the physical world has become a wasteland of darkened skies, polluted rivers and malicious forces whose purpose we can't comprehend. Many are even worse off since the government has not made education enough of a priority that its citizens can understand the real forces that control their future.

As in the Matrix, we are being harvested for our capacity as consumers, our ability to define new markets and spur competition to further pad the pockets of that shrinking percentile of human beings who have all the power, all the money and all the free time to enjoy what's left of the physical and social environment they have wrought, primarily by defining the ethos of our age.

We would be considerably poorer of spirit and shallow in our understanding of our shortcomings were it not for all that has been written about aesthetics, biology, history, jurisprudence, mathematics, neuroscience, physics, psychology, theology, and a score of other disciplines less well-known. All of these disciplines have helped shape humankind and many of us can say we would be substantially impoverished were it not for having read so broadly.

That said, there are many lessons that can be compressed into compact form, and, if delivered to a young student and periodically reinforced by a dedicated teacher at the right time during development, could probably have made that student significantly more patient, sanguine, and considerate to others in ways that could have a larger impact outside of the student's otherwise personal and parochial interests. Why is it that even the best teachers rarely convey to such important lessons?

The answer is simple. Those lessons are considered outside of their purview. Perhaps it was expected that our parents would provide those lessons, but who then would have taught our parents, and if the answer has the form of a regression, what stops it from regressing infinitely? Perhaps it is thought culture will serves to encourage such learning and that once it becomes widespread enough it is difficult not to acquire. But not all cultures have acquired this wisdom and some have lost it over time. The important thing is it can be taught and more importantly it can be learned if cultivated in a society that values it.

Once acquired, it is hard to imagine another way of engaging with one another and the world we share. It is at once common sense and uncommonly rare. The reason it is so rare is that until you experience engaging with someone possessed of such wisdom it is hard to imagine how powerful and life affirming it can be. It contradicts the natural way of interacting with the world defined narrowly as how natural selection has shaped our brains and bodies to survive. It requires a leap of faith that others will reciprocate and willingness to let down one's guard in advance of any guarantee that there will be any relaxation in hostility for the purpose of trying on a different approach to resolving conflict and pursuing mutual goals.

Some would say we are so much better off today owing to better nutrition, longer lives, improved sanitation, resistance to disease and pestilence, etc. and while true, none of these advantages confer wisdom or happiness. The fact that our brains evolved to interpret the world in ways that primarily served to improve our reproductive success does not imply that we can't see the world for what it is. Perhaps nature made a mistake when it made us smart enough to override instincts put in place by natural selection to promote our survival in the world in which we evolved.

5 Personal Assistants

Even in the best of worlds, with the most competent people acting on our behalf — people so well adjusted and naturally sanguine they can rise above their innate cognitive foibles, unaltered humans simply can't keep up with all the information vying for their attention. They need confederates to massage information into charts, talking points and simple summaries. When it comes to science and technology, even the most diligent and widely read scientists can hardly keep track of their own fields.

There are approximately 2.5 million new scientific papers are published each year. Lots of them are not relevant to the major social, political and environment issues of the day, but increasingly science and technology are defining the future and so someone has to sift out the manifests and papers that are relevant and bring them to the attention of policy makers.

A world leader has to keep track of what's going on in other countries, national politics, global social movements, terrorism, the consequences of new tariffs and trade negotiations, all the while keeping tabs on thousands or millions of constituents. Even with an army of aides and deputies it is impossible to organize and understand what's needed to make crucial decisions given all the factors that can influence the outcome.

Whatever their emotional and common-sense shortcomings, AI systems like IBM's Watson and DeepMind's AlphaGo are able to sift through mountains of raw data, read and summarize every scientific paper on a given subject in hours, simulate millions of scenarios in parallel and generate solutions to incredibly complex decision problems.

For such systems to produce solutions to problems we care about, they need to understand our hopes and dreams, but no simple polling approach will work to elicit our preferences since we don't know ourselves. If you want to know what someone thinks about global warming, unless they've thought carefully about the issues, sought out and read the relevant literature or credible summaries, then the best you can hope for is a second- or third-hand opinion.

Crowd sourcing doesn't work if people don't think for themselves. If a citizen doesn't have the necessary background, they are not likely to give you a week of their time to get them up to speed. Democracy requires an educated electorate and that takes time, money and patience. Ideally we'd like to model the perfect personal tutor — an infinitely patient machine that behaves like a close friend or colleague, always on the lookout for that perfect teaching moment to point out noteworthy items and relate them to current issues.

How can humans come to understand the emotional and social consequences of environmental or economic changes unless they engage with the issues? How else can humans form personal opinions without talking about the issues with one another and consulting experts representing different perspectives. If such learning and engagement was seamlessly integrated into fabric of everyday life it might not feel like being passively fed propaganda. If you could ask questions, seek out and contrast opposing viewpoints and engage in civil discourse to defend your opinions, school would not seem as onerous as your first exposure might have led you to believe.

AI enabled personal assistants could be trained to see around and under our biases, call out inconsistencies, inaccuracies and opinions shaped less by the facts and more by the amplified bias of crowds, paid political proselytizing and special-interest-sponsored infomercials. It will be a challenge to build systems that can be trusted to present balanced viewpoints, engage users transparently allowing them to make up their minds and seek out any sources they request, while allowing for the possibility they may fall prey to social pressure and choose with their cohort, but much less challenging than training a perfect human tutor for every person on the planet.

6 One Step at a Time

Technologically and socially speaking how do we transition from a world in which politicians, lobbyists, civil servants and government bureaucrats even at their best primarily serve narrowly-focused special interests and inadequately-informed and poorly-educated citizens to a world in which sophisticated cognitive prostheses enable us to make more-informed individual decisions and pool our narrow self interests and broader human aspirations to arrive at strategies that make the most of our limited resources to serve our needs, realize our ambitions and conserve and protect the environment to enhance our lives and those of our descendants.

The answer we offer involves building on existing technologies and infrastructure and providing incentives for entrepreneurs to add value to and extend these existing capabilities by developing services people

will pay for and at the same time provide technology that will serve to build the tools we need to realize the vision outlined in these pages. Many of the component technologies already exist in nascent form and will no doubt be improved upon to suit the needs of paying customers. Improved tools will leverage modern machine learning and AI technologies and their success in business will accelerate their development and deployment in the applications envisaged here.

We are building upon the solid foundation of a global communications network, vast information resources and powerful tools for extracting value from that information plus an incredible array of tools and services for sharing and collaborating broadly across social and geographic scales. The tools we are proposing here are natural extensions of technologies already in place to support individuals, businesses and municipalities. The extensions required will build upon existing support infrastructure for security, anonymity, sharing, verifiability and transparency allowing these capabilities to be combined to suit different purposes.

In addition, we imagine new tools for measuring affect and arousal to facilitate civil discourse and the free exchange of ideas, tools for individual and collective decision-making, tools for forging agreements and formulating contracts and policies that scale from small groups of neighbors to billions of widely distributed participants, tools to quantify the likelihood of uncertain outcomes and verify the truth or falsity of statements, and tools to assist in making value assessments and accurate predictions in support of long-term planning.

How do we make the transition from chatterbots that mimic therapists and remind us to take our medications to systems with deep knowledge of human nature and the ability to engage in prolonged dialogue aimed at helping you manage your emotions and make better decisions? In the following we will attempt to provide a partial answer to this question. A full answer would require addressing the social challenges of making such wholesale changes in society. Our reply is simply to note that the process of deployment we imagine would take place over decades giving individuals and institutions time to accommodate the changes gradually.

A full answer would also require a description of how and when AI systems could achieve a level of competence on a par with or exceeding the ability of trained therapists and facilitators to broker disputes and arrive at mutually agreeable compromises[2]. Fortunately, the approach outlined here does not require such advanced technology. We don't require systems that can read your mind or pass the Turing test. What we need are AI systems that monitor the biological markers of arousal and stress and intervene to assist in mitigating the personal and social consequences of engagement in such circumstances [9]. Basic interventional strategies are well studied and more personalized management techniques can be learned.

In addition to helping us avoid being hijacked by our sympathetic nervous system, we need technology that can answer questions, offer timely advice and help us deal with the vicissitudes of everyday life. That might sound like a tall order, but there is a large market for a range of such digital assistants, savvy consumers will ferret out the benevolent technologies, and the best will serve as the basis for the technology we have in mind. If this discussion seems facile, that is because the core technologies either exist or are rapidly evolving due to growing markets, highly motivated entrepreneurs and global investments in technology.

Exotic neural interfaces that are invasive (penetrate the skull) and chronic (remain in place and resist degradation) may significantly improve brain-compute communication. In the meantime, we can use natural language interfaces and exploit the most sophisticated, subtle and high-bandwidth method of communication ever conceived. One of the biggest technology challenges involves protecting user data. Fortunately, the *blockchain* technology used in cryptocurrencies like Bitcoin and Ethereum has been shown to scale to handle

[2]If deemed relevant, we can defer to the experts who attended the 2017 Beneficial AI Conference sponsored by the Future of Life Institute. This 2.5 day meeting included two special sessions including plenary talks and followed by their respective panels asking the questions "When can we expect human level AI?" and "When can we expect superintelligent systems?" Answers to the first question ranged from a few decades to centuries. Answers to the second question ranged from a few minutes after achieving AI level to never. A nuanced answer to either of the questions would require considerably more effort on the part of the authors and readers alike.

the security and accounting requirements for large companies like Nasdaq and small countries like Estonia, with broader adoption likely to follow.

7 Human Intelligence

Human AI is technology designed to behave similar to and interact naturally with human beings. Here we also assume that human AI systems are modeled after what we currently know about the human brain because we believe this will facilitate designing and building cognitive prostheses that will expand the capabilities of both natural and artificial humans.

Building the first human AI's designed to help natural humans realize their potential and ultimately collaborate to govern themselves will require engineering capabilities that, while they may seem incredibly ambitious, are within the scope of existing technology. We don't need human-level AI to begin with since we can bootstrap technology development.

To begin with we need to design systems capable of interacting with human beings on any topic that might come up in casual conversation. We propose to listen into billions of conversations between millions of pairs of humans engaging in everyday conversation and compile a large data set for training conversational systems. This approach has two potential shortcomings.

First, how do you convince humans to contribute their unedited and often intimate conversations to a database that might be co-opted for nefarious purposes? Technically, it is possible to anonymize such data so that it would be virtually impossible to trace the origins of any conversation back to its source. Socially, it will take time and care to convince people that their secrets are safe and anything they say won't be used in an improper way. Given the frequency and magnitude of just the reported data breaches, such reservations seem in order. However, there are technologies that are up to the task and as long as there are choices the market will provide solutions.

The second shortcoming is more of a technical challenge and basically comes down to a question of coverage and specific application. Wouldn't it require an impossibly large dataset to cover all of the possible conversations that might come up in subsequent interactions between humans and their human AI counterparts? The short answer is that modern neural network language systems are able to use such data sets to improvise by exploiting common patterns and re-contextualizing sentence fragments as required to suit individual circumstances.

The AI system doesn't need to have heard all possible conversations, just enough to recognize the common patterns of human discourse so it can adapt to whatever topic and conversational style it encounters. The AI system would also have to be trained how to recognize when it has misunderstood and gracefully recover and reestablish rapport, but this too is largely stylized and even humans routinely find themselves not understanding what their partner is trying to say and know when to give up and deftly redirect the conversation.

Our goal here is not build the best chatterbot but to learn how to converse naturally. Since we aim to build a combination of assistant, confidante and teacher, we want a system that is omnipresent but less intrusive than the other voices in our heads, available when we need it but easily dismissed when we prefer to be alone with our thoughts, and so versatile that we can depend on it for everything from reminders for your next dentist appointment to advice on how to get a good nights sleep prior to an important interview.

Engaging in the sort of unconstrained dialogue that one might hear around the proverbial water cooler, chance encounter on the bus or at the dinner table is one thing. Answering questions concerning society, politics, climate, etc. is obviously more challenging but in this case the interaction can be carefully scripted to account for what the human knows and believes, and what the human AI has to offer by way of technical explanation, fact checking and best practices for how to frame and explore a topic in order to arrive at a better understanding.

In principle, this sort of discussion could easily require the full range of capabilities of a highly trained and well-informed interlocutor. From a technical standpoint, the underlying problem is said to be AI complete, that is, requiring the full gamut of human intellect. While we expect that AI systems will achieve this in due course, we are interested here primarily in convincing the reader that the technologies required to implement the future we envision are relatively modest when compared to the aspirations of some AI enthusiasts.

Keep in mind that the systems we are considering here would continually be collecting dialogue to be used to improve their ability to address new topics and modes of discourse. Once again the resulting capability need only recognize the terminology and patterns of speech, since for the most part the same strategies for engagement will apply and only the facts, talking points and controversies will be new and require adaptation.

We are assuming that each individual will have his or her own dedicated virtual human AI personal assistant. While the centralized collection of anonymized data will protect the individual from any embarrassments or reprisals, our dedicated personal assistant will know us intimately requiring different strategies for protecting privacy. This intimacy and access to all of our personal conversations could be customized to suit the user's preferences, but our assistant will be able to assist us far better having been granted full access and there exist robust technologies and protocols that mitigate the risks of disclosure.

Such openness will be tested even further if we allow our assistants to record and make use of biological markers that provide insight into the factors governing their emotional and cognitive capacities relating to stress, alertness, anxiety, anger, disgust, etc. Knowing such information will be helpful not only in helping the assistant to understand what is motivating the user, but also in assessing the emotional state of the user to better understand the conditions under which utterances are produced and what that might tell us about the degree to which the user is invested in claims made and opinions offered.

The data corresponding to these biological markers, when aligned with the participants' individual contributions to a dialogue are doubly useful in learning about human discourse and inferring human preferences so as to better represent their intentions when collecting data to drive decision-making and generate policies. Collecting such data is also considerably more controversial for the same reason it is so revealing and diagnostically useful. Sandy Pentland has provided compelling technical arguments showing how such personal data can be suitably anonymized and how it can be applied to solve vexing problems in a wide range of social contexts.

The above discussion of how personal assistants might be engineered to be useful to both the individual and the collective depends on technologies we already have or likely will develop in the near future. The discussion suggests this can be accomplished without inventing full human-level AI and is agnostic on whether or when that might happen. However, we believe there is much we can can learn from studying human cognition. In particular, human AI systems employing cognitive strategies implemented on human-inspired cognitive architectures could help in developing advanced cognitive prostheses that benefit both humans and human AI systems.

The primate cortex consists of multiple functional areas arranged, roughly speaking, in a hierarchy with inputs and outputs at the bottom of the hierarchy and planning and executive control at the top. The lowest level in the hierarchy consists of the primary sensory and motor areas. As the names suggest, these areas integrate input from the sensory systems and generate output for the motor systems. The next level consists of association areas that integrate information from multiple modalities originating in the sensory and motor areas.

The activity of neurons within all of these areas can be thought of as providing two basic services: they serve as representations that we generically refer to as *concepts* and they perform computations that serve to construct and modify those concepts as new information becomes available[3]. The information produced by

[3]In the brain, all computations are carried out *in place* in contrast with conventional computers in which computations are carried

these computations is propagated within and between areas via *feedforward* and *feedback* connections.

Feedforward connections generally connect simpler concepts to more complex ones creating abstract thoughts. Feedback connections connect complex concepts to simpler concepts and in so doing provide expectations for what we have or might sense and suggestions for what we might or will do. *Recurrent* neural networks employ feedback connections to represent the dynamics of sequential processes that play out over time such as speech or movement.

Artificial neural network architectures are constructed by wiring together specialized layers using different types of connections. Each layer takes as input the output from other layers, computes as its output some function of that input and provides that output to other layers. These specialized layers include *convolutional* layers that take structured input, e.g., visual information from the retina, and provide structured output, e.g., a map representing the color registered at each location on the retina.

Attentional layers enable neural networks to spend more time on some parts of the input than others guided by a function that classifies parts of the input as being interesting or not. This classifier might be implemented as another layer trained by *reinforcement learning*, the same method used by DeepMind to train a neural network to excel at Go. In recent years, research in cognitive neuroscience has developed models of human attention and how we model the minds of others to infer what they know in order to facilitate sharing information.

The point here is not to give you a crash course in artificial neural networks but rather to point out that the field is continually inspired by new findings about the brain to create new classes of neural networks that capture aspects of biological computation. Inspiration can come from microscale studies involving neural circuits consisting of a few dozen neurons suggesting new ways of training artificial neural networks or from macroscale studies of networks consisting of components comprised of hundreds of thousands of neurons suggesting new architectures for automated planning and complex decision making.

This virtuous cycle of sharing ideas will serve to accelerate progress in both fields as the technology for recording from large numbers of neurons in awake behaving animals continues to improve. Indeed artificial neural networks are playing an important role in analyzing biological circuits by tracing the axons and dendrites of individual neurons in specially prepared samples of neural tissue and by inferring the function of specific circuits by recording from large numbers of neurons in animals performing specific tasks. There is advantage and precedent in creating our first intelligent beings in our image.

8 August Imagination

In the preface to a recent Edge conversation with the title, "Reality is an Activity of the Most August Imagination", drawn from a 1949 book by Wallace Stevens [12], Tim O'Reilly wrote:

> Our job is to imagine a better future, because if we can imagine it, we can create it. But it starts with that imagination. The future that we can imagine shouldn't be a dystopian vision of robots that are wiping us out, of climate change that is going to destroy our society. It should be a vision of how we will rise to the challenges that we face in the next century, that we will build an enduring civilization, and that we will build a world that is better for our children and grandchildren and great-grandchildren. It should be a vision that we will become one of those long-lasting species rather than a flash in the pan that wipes itself out because of its lack of foresight.

out in a central location and data that is stored elsewhere and has to be shuttled back and forth between the two locations in order to perform computations on that data. Since neurons both store information and perform computations on that information, the conventional distinction between computing elements and storage elements in computer science is more complicated in biological computation.

O'Reilly goes on to talk about what it is that humans collectively believe and how we attempt a construct vision about the way the world ought to work, mentioning the work of Yuval Noah Harari in which Harari traces the evolution of the collective beliefs that shape and limit how we behave to one another. Channeling Ezra Pound, O'Reilly suggests that "[w]e have to look at the world as it is and the challenges that are facing us, and we have to throw away the old stuck policies where this idea over here is somehow inescapably attached to this other idea. Just break it all apart and put it together in new ways, with fresh ideas and fresh approaches."

One of the ideas I believe we have to throw away is the conceit that by using technology to improve ourselves or replace institutions that don't scale or simply don't work, we are abandoning or denigrating our much vaunted humanity, when, in fact, human ingenuity created technology and it is as much an extension of our humanity as are the institutions we cherish and often forget are themselves technology that has emerged, evolved, been overthrown and superseded countless times in our history.

Educators, evangelists, poets, politicians and philosophers have been tinkering with our minds, shaping our thoughts and interceding in social discourse since we learned to use language. Ideas are like pieces of software that anyone with a modicum of seriousness or authority can deploy to reprogram your brain.

In this article, we are arguing that it is possible to gradually replace outdated institutions that don't scale and have caused at least as many problems as they have solved, using technology that does scale and enables us to channel our best selves to address our problems and govern our planet together.

The proposal here suggests we can replace the politicians and governments with technology that supports collaborative deliberation, enabling us to formulate our preferences informed by systems that collect, collate, extrapolate from and explain data, and can elicit and combine our preferences to formulate policies for governing the planet.

The systems we are referring to are based on artificial intelligence technology. Those that extrapolate data and integrate preferences are essentially the descendants of IBM's Watson and DeepMind's AlphGo — they are highly advanced calculators. Those that help shape the discourse between humans channel the best practices of behavioral sciences.

The path forward will likely take decades but it will proceed in small steps, leveraging existing technologies, working with individuals and small groups to improve the way we work together and gradually scaling to larger collectives providing plenty of time for institutions and society to adjust and tune best practices to suit local preferences.

Against the backdrop of these changes, AI technology will continue evolve and it will be necessary to accommodate enhanced (cybernetic) humans and, eventually, pure AI systems that are least as advanced and deserving of respect and a place at the table as are unenhanced humans. There will be time for us all to adjust and grow up together.

9 Acknowledgments

The author would like to thank the following for their feedback on an earlier draft: Ed Boyden, Yoshua Bengio, Robert Burton, Bryan Johnson, Christof Koch, Adam Marblestone, Sandy Pentland, Bart Selman, Rahul Sukthankar and Jaan Tallinn.

References

[1] Zoë Chance and Michael I. Norton. The what and why of self-deception. *Current Opinion in Psychology*, 6(Supplement C):104-107, 2015.

[2] Jared Diamond. *Collapse: How Societies Choose to Fail or Survive*. Penguin Books Limited, 2013.

[3] Yuval Noah Harari. *Sapiens: A Brief History of Humankind*. HarperCollins, 2015.

[4] D. Kahneman and A. Tversky. Choices, values, and frames. *American Psychologist*, 39(4):341-350, 1984.

[5] Daniel Kahneman and Amos Tversky. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5:207-232, 1973.

[6] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decisions under risk. *Econometrica*, 47:263-291, 1979.

[7] Shian-Ling Keng, Moria J. Smoski, and Clive J. Robins. Effects of mindfulness on psychological health: A review of empirical studies. *Clinical Psychology Review*, 31:1041-1056, 2011.

[8] W. Mischel. *The Marshmallow Test: Understanding Self-Control and How to Master It*. Transworld Publishers Limited, 2014.

[9] Alex (Sandy) Pentland. *Honest Signals: How They Shape Our World*. The MIT Press, 2008.

[10] Robert M. Sapolsky. *Behave: The Biology of Humans at Our Best and Worst*. Penguin Publishing Group, 2017.

[11] Megan K. Smith, Robert Trivers, and William von Hippel. Self-deception facilitates interpersonal persuasion. *Journal of Economic Psychology*, 2017.

[12] Wallace Stevens. *An Ordinary Evening in New Haven*. Connecticut Academy of Arts & Sciences, 1977.

[13] N. Thoma, B. Pilecki, and D. McKay. Contemporary cognitive behavior therapy: A review of theory, history, and evidence. *Psychodynamic Psychiatry*, 43(3):423-461, 2015.

[14] Robert Trivers. *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life*. Basic Books, New York, NY, 2011.