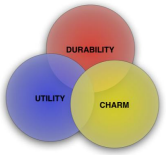
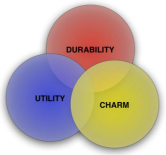
Digital Readiness: AI/ML Common Sense prevails?

Gregg Vesonder
Stevens Institute of Technology
http://vesonder.com



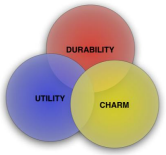
Three Talks

- ~~Digital Readiness: AI/ML, The thinking system quest.~~
  - ~~Artificial Intelligence and Machine Learning (AI/ML) have had a fascinating evolution from 1950 to the present. This talk sketches the main themes of AI and machine learning, tracing the evolution of the field since its beginning in the 1950s and explaining some of its main concepts. These eras are characterized as "from knowledge is power" to "data is king".~~
- ~~Digital Readiness: AI/ML, Finding a doing machine.~~
  - ~~In the last decade Machine Learning had a remarkable success record. We will review reasons for that success, review the technology, examine areas of need and explore what happened to the rest of AI, GOFAI (Good Old Fashion AI).~~
- Digital Readiness: AI/ML, Common Sense prevails?
  - Will there be another AI Winter? We will explore some clues to where the current AI/ML may reunite with GOFAI (Good Old Fashioned AI) and hopefully expand the utility of both. This will include extrapolating on the necessary melding of AI with engineering, particularly systems engineering.



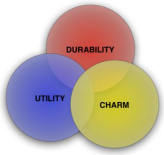
Winter is Coming?

- First Summer: Irrational Exuberance (1948 – 1966)
- First Winter (1967 – 1977)
- Second Summer: Knowledge is Power (1978 – 1987)
- Second Winter (1988 – 2011)
- Third Summer (2012 – ?)
- Why there might not be a third winter!



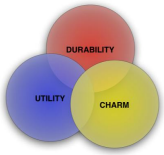
Roadmap

- Some preliminaries – the Turing Test
- ML + GOFAI
- Transparency
- Augmentation
- Safety in AI
- AI Self Awareness



Turing Test Needs

- Natural Language Processing (NLP)
- Knowledge Representation
- Automated Reasoning
- Machine Learning

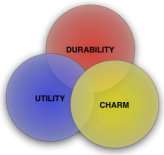


The Turing Test

- Turing paper "Computing Machinery and Intelligence"
- Behavioral test for intelligence
- Program has a conversation with a person for 5 min.
- Program passes test if it fools the person 30% of the time
- Turing predicted it would pass in 2000
  - Still not the case

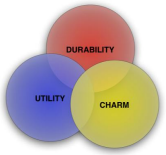
Stephen Wolfram: "A good Turing Test for me is When a bot can answer mot of my email"

As told by Russell and Norvig



Generalize

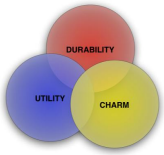
- "Current machine learning methods seem weak when they are required to generalize beyond the training distribution, which is what is often needed in practice." Yoshua Benigo



ML and GOFAI: Neuro[Symbolic]

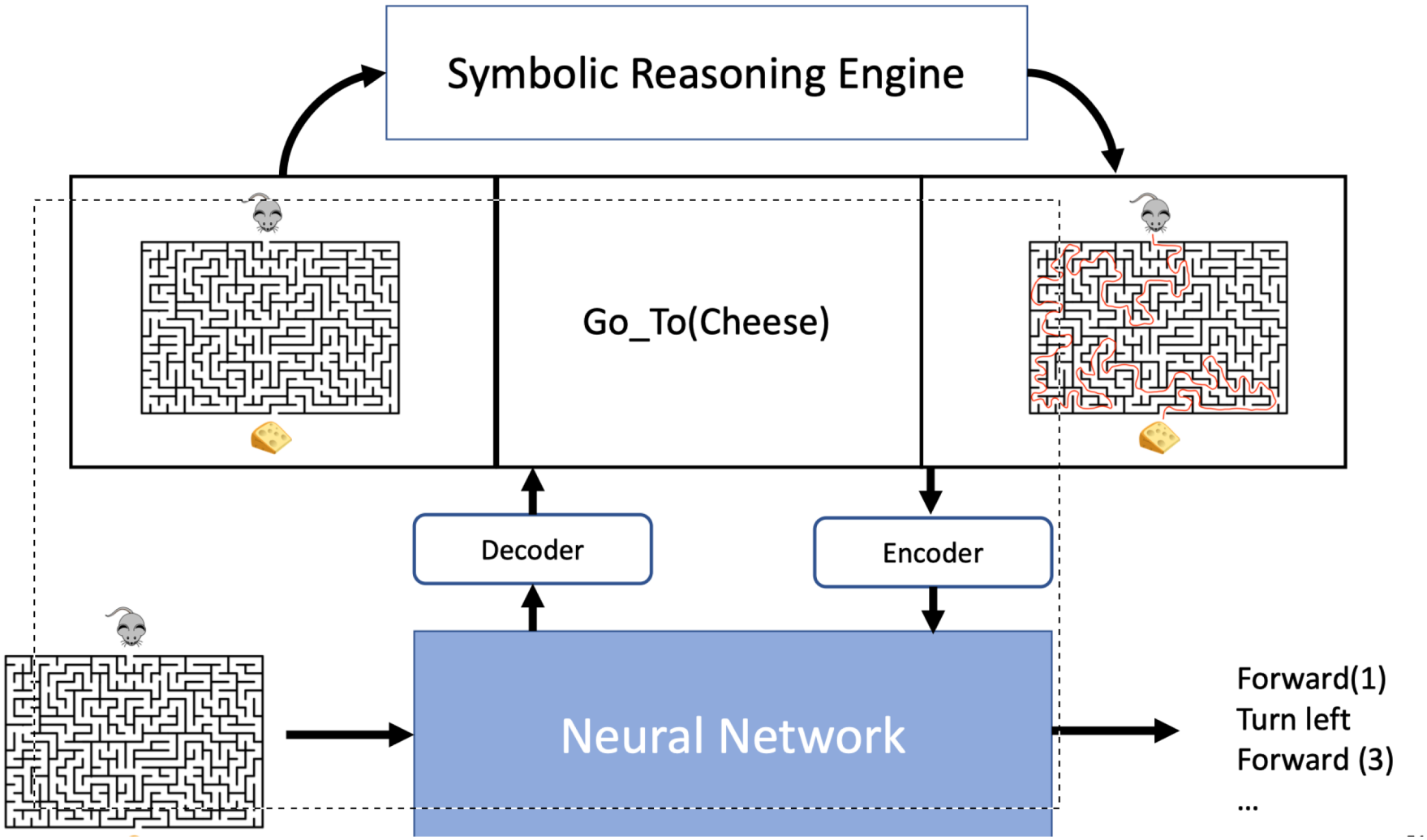
- Imbed true symbolic reasoning inside a neural engine
  - Enable super-human and super-neuro combinatorial reasoning
  - For deliberative, Type 2 reasoning
  - Rare in ordinary animal life
  - Common in "business AI"
- Interface at the Attention Schema
  - Internal model of the system's state of attention

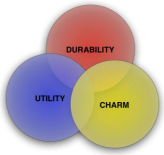
Not the same as attention itself!



Henry's favorite Neuro[Symbolic]

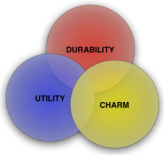
- Proposal: When attention to concepts is very high, they are decoded to symbolic entities in an attention schema
- Appearance of a goal in the attention schema signals that deliberative symbolic reasoning should be initiated
  - *I'm not just perceiving or remembering, I'm consciously thinking!*
- Efficient combinational search can then be performed over the entities in the attention schema


Symbolic Reasoning Engine
Go_To(Cheese)
Decoder
Encoder
Neural Network
Forward(1)
Turn left
Forward (3)
...



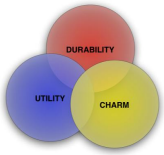
Some Early Predictions

- In 1965 Herbert Simon said, "Machines will be capable, within 20 years, of doing any work a man can do."
- Two years later, MIT researcher Marvin Minsky predicted, "Within a generation ... the problem of creating 'artificial intelligence' will substantially be solved."



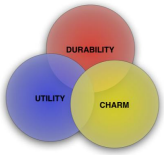
Philosophers and AI

- Weak AI – machines act as if they were intelligent
  - Dartmouth workshop 1955 asserted this would be the case
- Strong AI hypothesis- machines are actually thinking
- Engineering AI – quest for the best AI agent on a given platform/architecture
- Turing: a machine can never do X
  - Kind, resourceful, beautiful, friendly, have initiative, have a sense of humor, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make someone fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behavior as man, do something really new.



Desired – Deep Understanding: Common Sense

- Capacity to look at any scenario and address questions such as a journalist might ask: who, what, where, why, when, and how
- Needs commonsense knowledge, explicit reasoning, explicit cognitive models of the world
- John McCarthy "One will be able to assume that [the proposed system] will have available to it a fairly wide class of immediate logical consequences of anything it is told and its previous knowledge. This property is expected to have much in common with what makes us describe certain humans as having common sense"

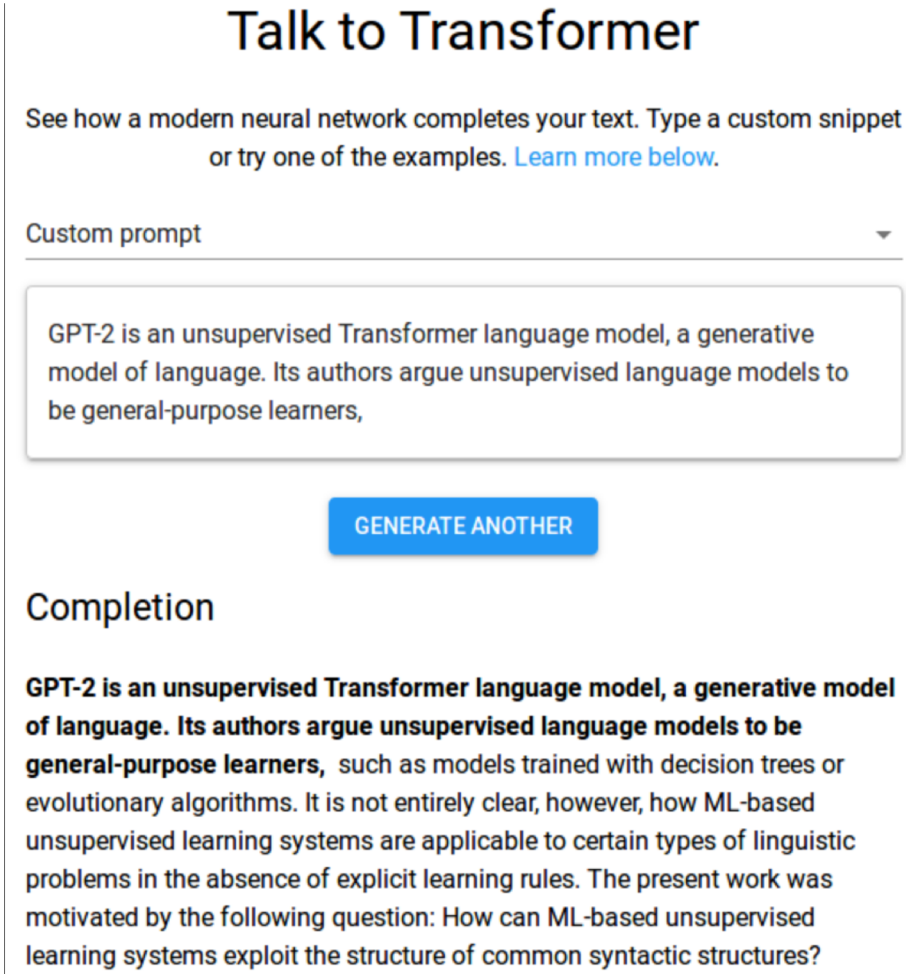


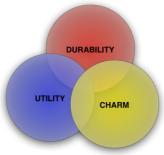
The New Wave

- Open AI - Elon Musk
- Deep Mind - Google

GPT-2 (OpenAI)

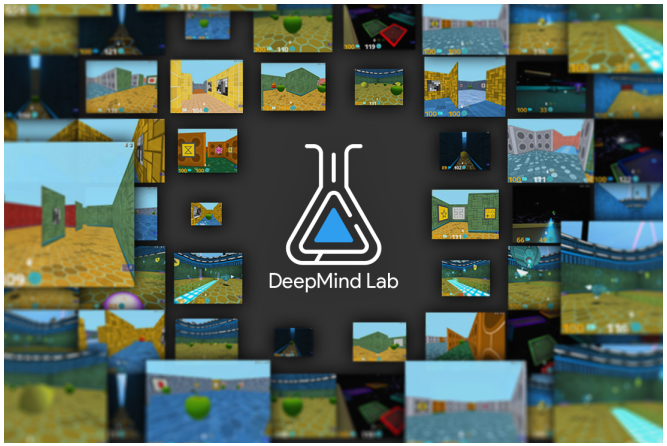
- Generative Pretrained Transformer-2
- Trained on WebText – 8 million documents
- Constrained due to spam fears

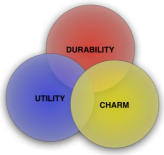




DeepMind

- "we aim to build advanced AI - sometimes known as Artificial General Intelligence (AGI)"
- DeepMind Lab – create realistic virtual worlds for AI research to explore General intelligence, Quake III arena levels, minecraft
- Platform (simulation) for studying intelligence
- Competition MineRL





Even Now

- Dangerous applications are enabled by third summer AI
- Threats most prominent in the news might not be the worst!
- Three examples:
  - Face recognition
  - Fake news
  - Autonomous weapons

Keeping your face private

Facebook to Pay $550M to Settle Class Action Case Over Facial Recognition

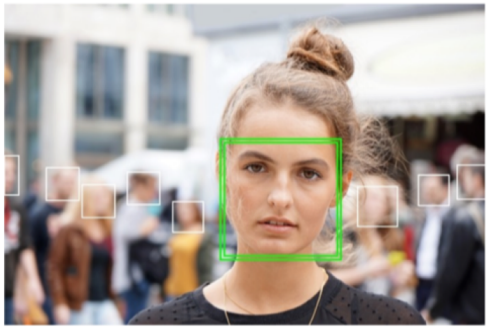
Author:
Elizabeth Montalbano

January 30, 2020
/ 7:05 am

2 minute read

Write a comment

Share this article:



The settlement in a case over the social network's Tag Suggestions feature is the latest financial blow the company has taken over its handling of user privacy.

San Francisco Bans Facial Recognition Technology

Attendees interacting with a facial recognition demonstration at this year's CES in Las Vegas. Joe Buglewicz for The New York Times

By Kate Conger, Richard Fausset and Serge F. Kovaleski

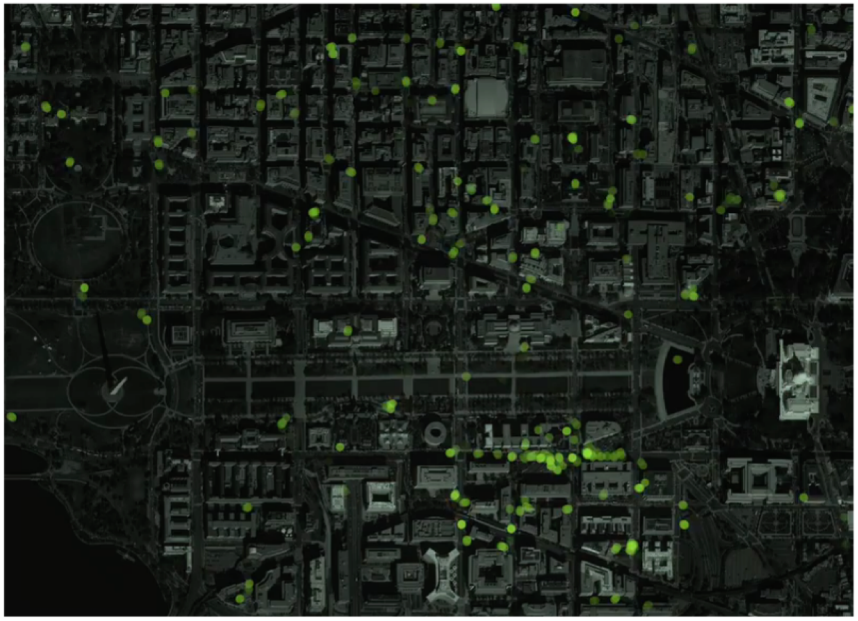
May 14, 2019

362

SAN FRANCISCO — San Francisco, long at the heart of the technology revolution, took a stand against potential abuse on Tuesday by banning the use of facial recognition software by the police and other agencies.

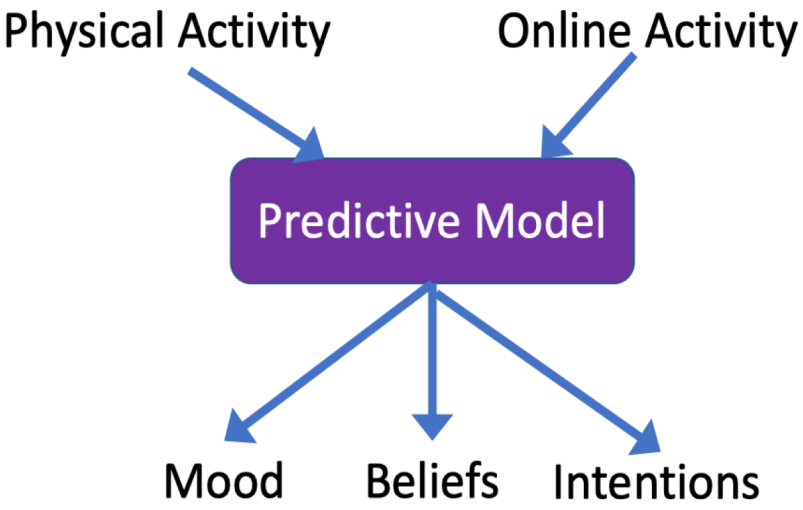
Keeping your face private - 2

- Face Recognition == Threat
- But your mobile phone apps are giving your location away – no AI needed!
- A. Sadilek, H. Kautz & J. P. Bigham (2012), Finding Your Friends and Following Them to Where You Are
- L. Liao, D. J. Patterson, D. Fox, and H. Kautz (2007), Learning and Inferring Transportation Routines



New York Times, "Twelve Million Phones, One Dataset, Zero Privacy", Dec. 19, 2019

Greater threat: Keeping your mind private



But this is about BAD…

Greater threat: Keeping your mind

Xinjiang Re-Education Camp

- Totalitarian states are inefficient when they target entire demographic groups for repression
  - Cost
  - Radicalization
  - International pressure
- AI enables precise micro-targeting of wrong-thinkers

2024

George Orwell

In the news: Fake news

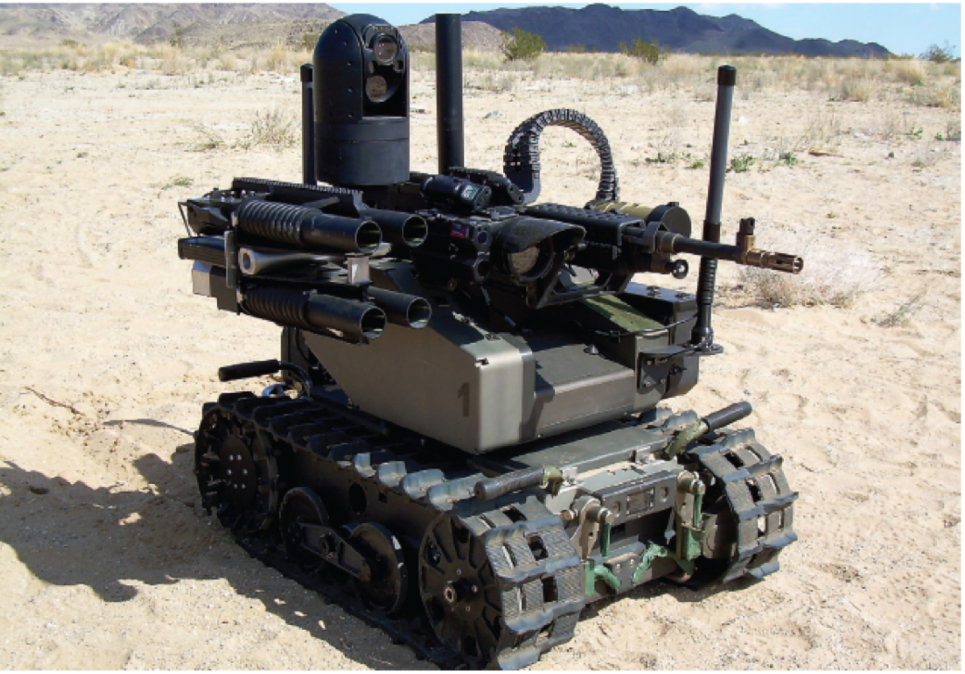
Artificial intelligence (AI)

New AI fake text generator may be too dangerous to release, say creators

The Elon Musk-backed nonprofit company OpenAI declines to release research publicly for fear of misuse

▲ The AI wrote a new passage of fiction set in China after being fed the opening line of Nineteen Eighty-Four by George Orwell (pictured). Photograph: Mondadori/Getty Images

In the news: Autonomous Weapons



Modular Advanced Armed Robotic System
Quentic North America

Defense Innovation Board AI Ethical Guidelines

- Human beings should exercise appropriate levels of judgment and remain responsible for the development, deployment, use and outcomes of DOD AI systems.
- Does this mean human-in the loop?

Ethical Military Use of AI != Human in the Loop

Larry Lewis

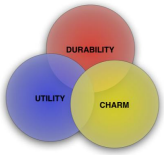
- Senior Advisor for the State Department on Civilian Protection, Obama administration
- Member US Delegation for UN Deliberations on Lethal Autonomous Weapons Systems
- Artificial intelligence may make weapons systems and the future of war relatively less risky for civilians than it is today

Killer robots reconsidered: Could AI weapons actually cut collateral damage?

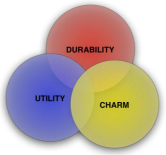
By Larry Lewis, January 10, 2020

The Bulletin of the Atomic Scientists, 10 Jan 2020

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY



Addressing the Future



Issues: Ethics of Computing

- People might lose their jobs to automation
- People might have too much or too little leisure time
- People might lose their sense of being unique
- AI systems might be used to undesirable ends
- The use of AI systems might result in a lack of accountability
- The success of AI might mean the end of the human race

Russell and Norvig

AI and Systems Engineering

Augmented AI
Systems Engineering

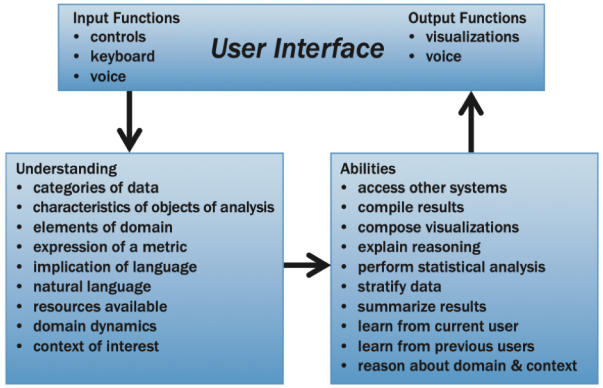


Figure 4. Understanding and abilities needed to augment Systems Engineering

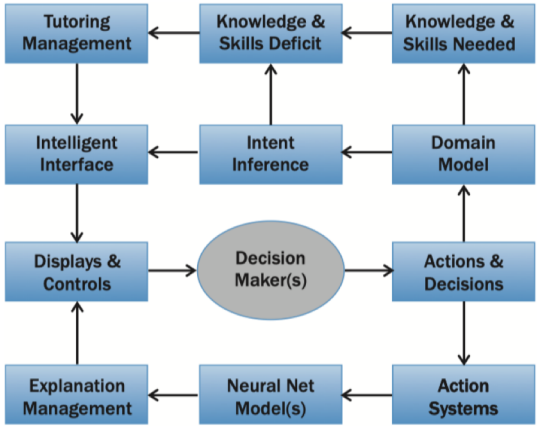
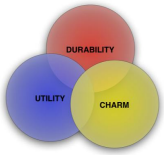


Figure 2. Overall architecture of augmented intelligence

William B. Rouse,"AI as system engineering augmented intelligence for systems engineers INCOSE, March 2020

AI and Software Engineering



Carnegie Mellon University
Software Engineering Institute

Designing Ethical AI Experiences: Checklist and Agreement

USE THIS DOCUMENT TO GUIDE THE DEVELOPMENT of accountable, de-risked, respectful, secure, honest, and usable artificial intelligence (AI) systems with a diverse team aligned on shared ethics. An initial version of this document was presented with the paper *Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development* by Carol Smith, available at https://arxiv.org/abs/1910.03515.

We will design our AI system with the following in mind:

- ☐ Designated humans have the ultimate responsibility for all decisions and outcomes:
  - Responsibilities are explicitly defined between the AI system and human(s), and how they are shared.
  - Human responsibility will be preserved for final decisions that affect a person's life, quality of life, health, or reputation.
  - Humans are always able to monitor, control, and deactivate systems.
- ☐ Significant decisions made by the AI system will be
  - explained
  - able to be overridden
  - appealable and reversible

We work to speculatively identify the full range of risks and benefits:

- ☐ Harmful, malicious use and consequences, as well as good, beneficial use and consequences
- ☐ We will be cognizant and exhaustively research unintended consequences.

We will create plans for the misuse/abuse of the AI system, including the following:

- ☐ communication plans to share pertinent information with all affected people
- ☐ mitigation plans for managing the identified speculative risks

We value respect and security:

- ☐ incorporating our values of humanity, ethics, equity, fairness, accessibility, diversity, and inclusion
- ☐ respecting privacy and data rights (Only necessary data will be collected.)
- ☐ providing understandable security methods
- ☐ making the AI system robust, valid, and reliable

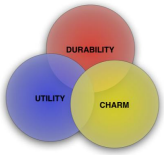
We value transparency with the goal of engendering trust:

- ☐ The purpose, limitations, and biases of the AI system are explained in plain language.
- ☐ Data sources have unambiguous respected sources, and biases are known and explicitly stated.
- ☐ Algorithms and models are appropriate and verifiable.
- ☐ Confidence and context are presented for humans to base decisions on.
- ☐ Transparent justification for recommendations and outcomes is provided.
- ☐ Straightforward and interpretable monitoring systems are provided.

We value honesty and usability:

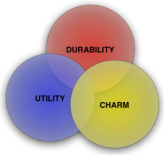
- ☐ Humans can easily discern when they are interacting with the AI system vs. a human.
- ☐ Humans can easily discern when and why the AI system is taking action and/or making decisions.
- ☐ Improvements will be made regularly to meet human needs and technical standards.

Team Signatures and Date



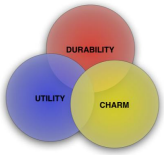
Example: 6 Levels of Vehicle Autonomy

- 0 – no driving automation
- 1 – driver assistance (adaptive cruise control)
- 2 - partial driving automation (steering and acceleration with human monitoring)
- 3 – conditional driving automation environmental detection, accelerating around slow moving vehicle (human monitoring)
- 4 – high driving automation vehicle can intervene if something goes wrong – human can still take control
- 5 – full driving automation effectively no pedals or wheels, no geo-fencing



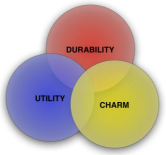
Robust AI

- Gary Marcus
  - Apply what it knows to a wide range of problems in a systematic and reliable way
  - Synthesizing knowledge from a variety of sources so that it can reason flexibly and dynamically
  - Transferring what it knows from one contextto another
- "Quite simply if we cannot count on our AI to behave reliably we should not trust it"



NSF and Robust Intelligence

Robust Intelligence (RI) encompasses foundational computational research needed to understand and develop systems that can sense, learn, reason, communicate, and act in the world; exhibit flexibility, resourcefulness, creativity, real-time responsiveness and long-term *reflection*; use a variety of representation or reasoning approaches; and demonstrate competence in complex environments and social contexts.

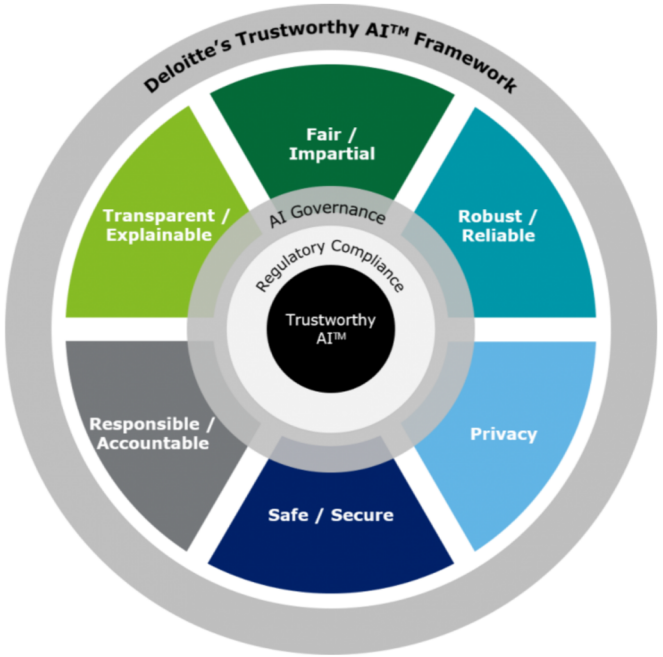


NSF and Institute Track

- Proposals for the **Institute** track must have a principal focus in one or more of the following themes:
  - Trustworthy AI;
  - Foundations of Machine Learning;
  - AI-Driven Innovation in Agriculture and the Food System;
  - AI-Augmented Learning;
  - AI for Accelerating Molecular Synthesis and Manufacturing; and
  - AI for Discovery in Physics.

Trustworthy AI Framework

- Fair, not biased
- Transparent and explainable
- Responsible and accountable
- Robust and reliable
- Respectful of privacy
- Safe and secure



MIT Tech Review, March 2020

Explainable AI

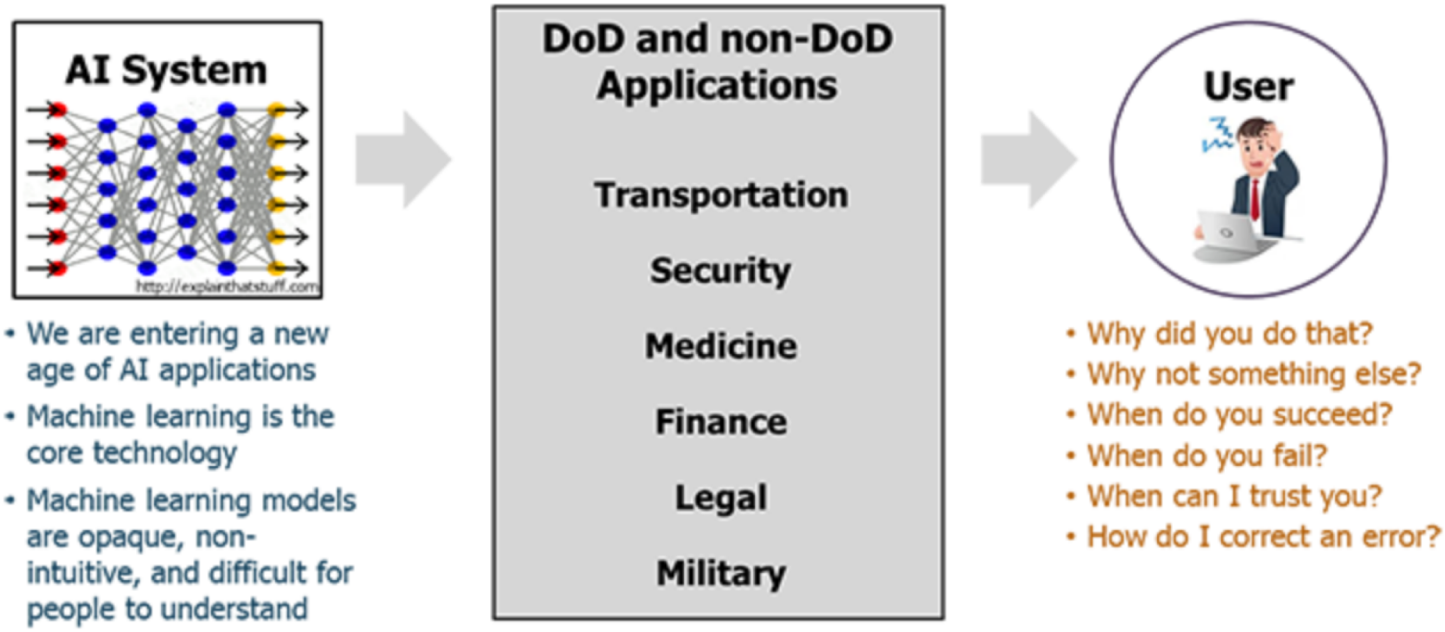


Figure 1. The Need for Explainable AI

Dr. Matt Turek, Explainable AI, XAI

Explainable AI -2

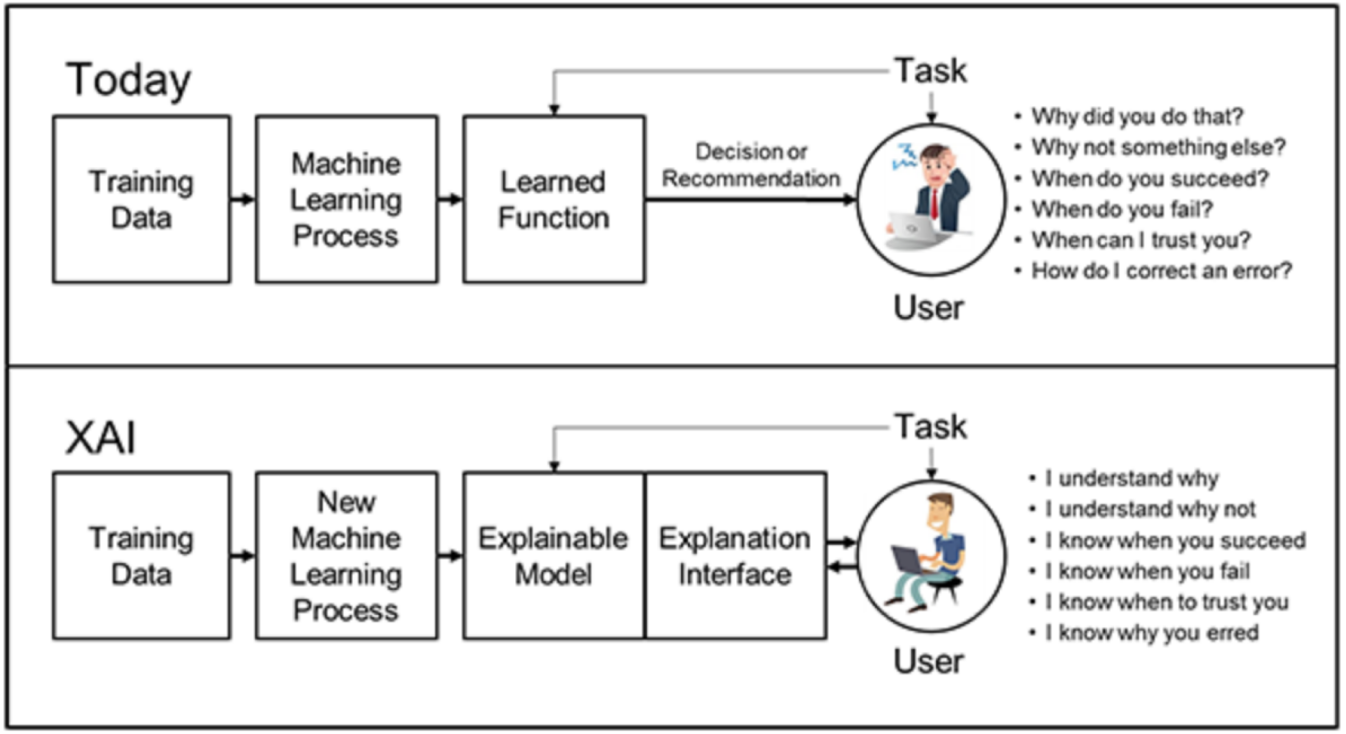
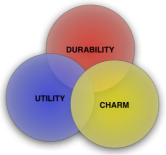
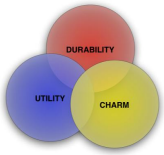


Figure 2. XAI Concept



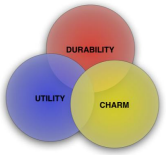
Asimov's Laws of Robotics

- 0. A robot may not injure a humanity or, through inaction, allow humanity to come to harm.
- 1. A robot may not injure a human being or, through inaction, allow a human being to come to harm, except where that would conflict with the Zeroth Law.
  - (old 1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.)
- 2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
- 3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.



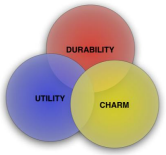
Canadian Government Montreal Declaration

- Montreal Declaration for a Responsible Development of AI (2018)
  - Develop an ethical framework for the development and deployment of AI
  - Guide the digital transition so everyone benefits from this technological revolution
  - Open a national and international forum for discussion to collectively achieve equitable, inclusive and ecologically sustainable AI Development



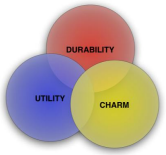
Montreal Declaration Principles

- Well Being – AIS must help individuals lead better lives, not become a source of ill-being
- Autonomy – AIS must help individuals to fulfil their own moral objectives and their conception of a life worth living
- Protection of Privacy and Intimacy – persona spaces in which people are not subjected to surveillance or digital evaluation must be protected from the intrusion of AIS
- Solidarity – the development o fAIS must be compatible with maintaining the bonds of solidarity among people and generations
- Participation – AIS must meet intelligibility, justifiability and accessibility criteria and must be subjected to democratic, scrutiny, debate and control



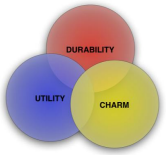
Montreal Declaration Principles -2

- Equity – The development and use of AIS must contribute to the creation of a just and equitable society
- Diversity Inclusion – The development and use of AIS must be compatible with maintaining social and cultural diversity and must not restrict the scope of lifestyle choices or personal experiences
- Prudence – every person involved in AIS development must exercise caution by anticipating as far as possible the adverse consequences of AIS use and by taking the appropriate measures to avoid them
- Responsibility – The development and use of AIS must not contribute to lessening the responsibility of human beings when decisions must be made



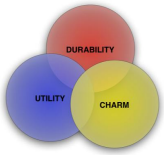
Montreal Declaration Principles - 3

- Sustainable Development – development and use of AIS must be carried out so as to ensure a strong environmental sustainability of the planet



Singularity

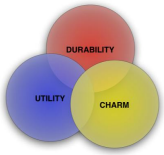
- First Vernor Vinge - https://frc.ri.cmu.edu/~hpm/book98/com.ch1/vinge.singularity.html



Singularity - 2

- *Computers that are "awake" and superhumanly intelligent may be developed.* (To date, there has been much controversy as to whether we can create human equivalence in a machine. But if the answer is "yes," then there is little doubt that more intelligent beings can be constructed shortly thereafter.)
- *Large computer networks and their associated users may "wake up" as superhumanly intelligent entities.*
- *Computer/human interfaces may become so intimate that users may reasonably be considered superhumanly intelligent.*
- *Biological science may provide means to improve natural human intellect.*

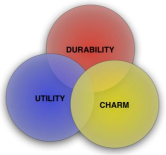
I'll be surprised if this event occurs before 2005 or after 2030.



Singularity - Kurzweil

- Continued exponential growth in computational power
- Sees it as humans transcending biology
- Uploading a specific brain with every process intact on a substantially powerful substrate
- A transition away from biological roots

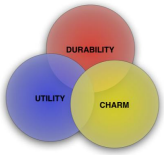
Thanks wikipedia



Richard Sutton on AI

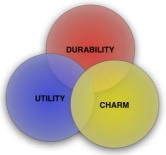
A human-level AI will be a profound scientific achievement which will happen by 2030(25% probability), 2040(40% probability), or never(10% probability)

Reinforcement Learning
Synced global AI newsletter



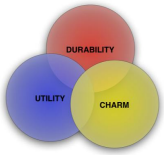
Stephen Hawking on AI

In short, the advent of super-intelligent AI would be either the best or the worst thing ever to happen to humanity. The real risk with AI isn't malice, but competence. A super-intelligent AI will be extremely good at accomplishing its goals, and if those goals aren't aligned with ours we're in trouble. You're probably not an evil ant-hater who steps on ants out of malice, but if you're in charge of a hydroelectric green-energy project and there's an anthill in the region to be flooded, too bad for the ants. Let's not place humanity in the position of those ants.



Elon Musk on AI

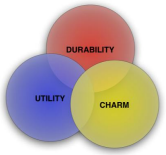
"AI will be the best or worst thing ever for humanity."



AI Sci Fi Books

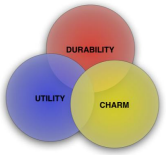
- Neuromancer – William Gibson
- Peripheral – William Gibson (simulations)
- Lifecycle of Software Objects – Ted Chiang
- I, Robot – Isaac Asimov

More AI resources will be available in a few days at vesonder.com



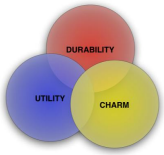
Winter is Coming?

- First Summer: Irrational Exuberance (1948 – 1966)
- First Winter (1967 – 1977)
- Second Summer: Knowledge is Power (1978 – 1987)
- Second Winter (1988 – 2011)
- Third Summer (2012 – ?)
- Why there might not be a third winter!



Three Talks

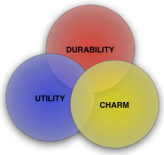
- ~~Digital Readiness: AI/ML, The thinking system quest.~~
  - ~~Artificial Intelligence and Machine Learning (AI/ML) have had a fascinating evolution from 1950 to the present. This talk sketches the main themes of AI and machine learning, tracing the evolution of the field since its beginning in the 1950s and explaining some of its main concepts. These eras are characterized as "from knowledge is power" to "data is king".~~
- ~~Digital Readiness: AI/ML, Finding a doing machine.~~
  - ~~In the last decade Machine Learning had a remarkable success record. We will review reasons for that success, review the technology, examine areas of need and explore what happened to the rest of AI, GOFAI (Good Old Fashion AI).~~
- ~~Digital Readiness: AI/ML, Common Sense prevails?~~
  - ~~Will there be another AI Winter? We will explore some clues to where the current AI/ML may reunite with GOFAI (Good Old Fashioned AI) and hopefully expand the utility of both. This will include extrapolating on the necessary melding of AI with engineering, particularly systems engineering.~~



This Summer

- Digital Readiness: Age of Digital Engineering, Dr. Jon Wade
- Digital Readiness: Drivers, Challenges, Opportunities, Mr. Troy Peterson
- Digital Readiness: Surrogate Pilot Experiments, Dr. Mark R. Blackburn
- The World of Data, Dr. Carlo Lipizzi
- Data and the World: State of Practice, Dr. Carlo, Lipizzi
- Data for the Upcoming World: Horizon Scanning, Dr. Carlo Lipizzi
- The Thinking Systems Quest, Dr. Gregg Vesonder
- Finding a Doing Machine, Dr. Gregg Vesonder
- Common Sense Prevails, Dr. Gregg Vesonder

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY



Thank You

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY