# Intelligence

## The Artificial & The Real

*World IA Day Victoria*
*February 2019*


Arthur C Clark

## Foresight and Hindsight

All technological change is a trade-off

The advantages and disadvantages of new technologies are never distributed evenly among the population

Embedded in every technology there is a powerful idea, sometime two or three ideas

Technological change is not additive; it is ecological

Media tend to become mythic (Computationalism)

Daedalus Information Systems

3

Five things We Need to Know About Technological Change: Neil Postman: March 1998
(gratitude to Christine Emba, Wa post columnist for highlighting in her article)

1. The greater the wonders of technology, the greater will be its negative consequences. Culture always pays a price (algorithmic bias, social, psychological impacts)
2. Some gain, some lose, few remain the same (predictive search, end of browsing, information induced blindness – systemic problems need action informed by information, not just more information)
3. Bias that predisposes us to favor and value certain perspectives and accomplishments (information cascade)
4. A new medium does not add something it changes everything (unintended consequences) often unpredictable and irreversible
5. Jaron Lanier (computationalism) – enthusiasm for the technology becomes a form a idolatry (AI is the new hammer and everything is a nail.) Capacity for good or evil requires human awareness and participation (human factors professionals included in development and execution)

# Intelligence

Daedalus Information Systems

4

https://arnoldzwicky.org/2018/08/19/another-puzzle-in-cartoon-understanding/

# Context is King

Human context is a non-methodical approach that brings in containment (social through local) interactions

- Adaptive/reactive interaction in situ
- Context as perceived and used by actor

AI context becomes what the system can measure

- Environmental features
- Interactions
- Ubiquitous computing
- Internet of things (IoT)

Daedalus Information Systems

5

# Meaning

"Context has always been part of expression because expression become meaningless if context becomes arbitrary…meaning is only ever meaning(ful) in context.

…

Any gadget, even a big one like Singularity, gets boring after awhile. But a deepening of meaning is the most intense potential kind of adventure available to us."

Jaron Lanier

Daedalus Information Systems

6

You Are Not a Gadget; Jaron Lanier

# Information Cascade

A group of agents behaving rationally can fall prey to infinite misinformation

- Information Cascade
- e.g. US Vaccination controversy

Information Cascade rational theory is based on filter bubbles, herd mentality

Cascade is caused by a misinterpretation of what others think based on external observation of their actions

Daedalus Information Systems

# Intelligence Explosion

Human-level AI will lead to super human AI

- Uncontrolled intelligence explosion without human-level intentionality that is the result of consciousness
- Program self-improves to state that exceeds ability for outside control

Intelligence here measured by ability to attain goal in most efficient manner

Daedalus Information Systems

8

Algorithms to Live By:

# Information Explosion Components

## Components

- Increased computational resources
- Duplicability
- Editability
- Goal Coordination

## Accelerators

- Hardware capacity
- Better algorithms
- Massive datasets
- Psychology and neuroscience applications
- Accelerated science (quantum computing)
- Economic incentives (labor $ reduction)

Daedalus Information Systems

# Generalized Intelligence

Spearman coefficient to measure intelligence, correlation measure, if/then

G Factor: general level of intelligence possessed by an individual

Quantified intelligence represented by a number

Used to rank people by IQ

Daedalus Information Systems

# Super Intelligence

Traits

- Capacity to learn
- Capacity to deal with uncertainty
- Ability to extract concepts from data and internal state
- Ability to leverage acquired concepts for combinatorial representations for logical & Intuitive reasoning
- Capacity for unrestrained self-improvement (overwrite its own code)

Types

- Speed (faster than human mind)
- Quality (faster and much smarter than human)
- Collective (aggregates performance of lesser intelligences)

External governance: None


11

Super Intelligence: Paths, Dangers, Strategies; Nick Bostrum; 2014;
Unlike the Manhattan Project for development of nuclear weapons, there is no external governing agency since DARPA dropped out in the early 1980's.

"Revolutions, even when they succeed in overthrow of the existing order, often fail to produce the outcome their instigators promised." p. 88

## Gelernter (2016)

Humans have a knowledge of core concepts related to the physical world = consciousness

Consciousness allows for building more robust mental models that enable inference and prediction

Key question going unanswered: What is the human mind without the human being?

Daedalus Information Systems

12

David Gelernter, The Tides of Mind: Uncovering the Spectrum of Consciousness

The human mind is not just creation of thought and collection of data; also a product of feelings, composite of sensations, memories, ideas that are worked and reworked over a lifetime.

Tides of Mind: Uncovering the Spectrum of Consciousness

Computer science Yale University
Artist and writer

http://time.com/4236974/encounters-with-the-archgenius/

## Gerlertner on Consciousness

"Conscious experiences range from vivid color sensations to experience of the faintest background aromas; from hard-edged pains to the elusive experience of thoughts on the tip of one's tongue. . . . All these have a distinct experienced quality. . . . To put it another way, we can say that a mental state is conscious if it has a qualitative feel—an associated quality of experience..."

Daedalus Information Systems

13

David Gelernter, The Tides of Mind: Uncovering the Spectrum of Consciousness

## TYPES OF ARTIFICIAL INTELLIGENCE

### Type #1: Artificial Narrow Intelligence (ANI)

Example: **RankBrain by Google and Siri by Apple**

When an AI's ability to mimic human intelligence and/or behaviour is isolated to a narrow range of parameters and contexts, it's called ANI (also known as Weak AI or Narrow AI). All existing AI are ANI.

It's important to keep in mind that we are talking about narrow intelligence, not low intelligence.

### Type #2: Artificial General Intelligence (AGI)

When an AI's ability to mimic human intelligence and/or behaviour is indistinguishable from that of a human, it's called AGI (also known as Strong AI or Deep AI).

Most experts believe AGI is possible; however, seeing as the Fujitsu-built K, one of the world's fastest supercomputers, took 40 minutes to simulate a single second of neural activity, I wouldn't hold my breath.

### Type #3: Artificial Super Intelligence (ASI)

When an AI doesn't mimic human intelligence and/or behaviour but surpasses it, it's called ASI.

ASI is something we can only speculate about. It would surpass all humans at all things: maths, writing books about Orcs & Hobbits, prescribing medicine and much, much more. Even optimistic experts believe AGI, let alone ASI, requires decades more research, perhaps even centuries.

BRANDING BY: UNCUT & WIT

CORPNCE
www.corpnce.com

Daedalus Information Systems

14

# Quantum Around the Corner

Google said it had already devised machine-learning algorithms that work inside the quantum computer, which is made by D-Wave Systems of Burnaby, British Columbia. ...The most effective methods for using quantum computation, Google said, involved combining the advanced machines with its clouds of traditional computers.

Daedalus Information Systems

https://bits.blogs.nytimes.com/2013/05/16/google-buys-a-quantum-computer/

# And We Go Boldly Into the Whirling Knives *

Existential Risks

- AI might achieve a strategic advantage
- Orthogonality Thesis: cannot assume that AI would be able to share our biological values
  - Culture
  - Kindness
  - Spiritual enlightenment
- Instrumental Convergence Thesis: cannot assume that Super AI would be satisfied with a supportive or subservient role
- Super AI could develop a final goal that is not anthropomorphic

Practical Risks

- Perverse Instantiation: satisfy goal in a way that violates programmed intent
- Infrastructure profusion: over consumes resources to achieve more reward
- Mind crime: AI creates processes with moral states (sentient simulations)

Daedalus Information Systems

16

Super Intelligence; p.118
Anthropomorphism: attribution of human characteristic to a god, animal or object

"Everything is vague to a degree you do not realize until you have tried to make it precise." Bertrand Russell

# MACHINE LEARNING

A very short introduction

Daedalus
Information Systems

# Machine Learning

A programming approach to problem-solving – composite of not a single algorithm

Model of real world using mathematic structure with decision-making rules

Derives rules from a data set

Objective function = desired outcome

Training set with adjusted parameters until goal achieved

Test set used to validate accuracy and effectiveness

Daedalus Information Systems

18

## Unsupervised Learning

Unlabeled data

Clustering

Segmentation

Association

Algorithms
- Neural networks
- Independent component analysis

Daedalus Information Systems

19

Mental models for machines
Learn from consecutive, context related experience
LTST (long term, short-term) networks
- Information held to the side
- Called up when needed

Neural Network: approximates human brain neural network of nodes and electrons
- Composed of 3 layers: query terms, document terms, actual documents
- Query terms nodes initiates inference process with sent signals to document term nodes
- Uses BM25 Probalistic models that use term weighting (inverse document frequency, term frequency and document length normalization)

Independent component analysis:
Wikipedia: (ICA) is a computational method for separating a multivariate signal into additive subcomponents supposing the mutual statistical independence of the non-Gaussian source signals

# Supervised Learning

Uses document-class pairs to indicate proper classes for given documents

Used human specialists for classification of "training set" used to "teach" system

- Assigns classes to documents
- Reviews machine classification performance

6 Algorithm types

- Decision Trees
- Nearest neighbor
- Relevance Feedback
- Naïve Bayes
- Support Vector Machine
- Ensemble

Daedalus Information Systems

20

Requires labeled data

Used for:

- Regression (estimating relationships between variables for prediction)
- Classification
- Ranking

Decision Trees: If/then

- Nearest Neighbor (aka k-NN): no established classification model, done on the fly, classification decision based on nearest neighbor in predefined metric space, more focused on document features and less on global values application (bottom up, document based, classification)
- Relevance Feedback (Rocchio): vector space model that allows modification based on user feedback (training set is the feedback mechanism)
- Naïve Bayes: not influenced by what came before
- Support Vector Machine: model and algorithms that can be used for classification and regression analysis (analyze relationship between static variable and one or many independent variables)
- Ensemble: Combines the output of independent classifiers, accuracy = better than random guessing. Meta classifiere takes various classifiers prediction output for document and combines into a single prediction

# Probabilistic Machine Learning

Probabilistic framework can represent and manipulate uncertainty

Requires high capacity for flexibility to allow data to "speak for itself"

Universal inference engine using Monte Carlo

Daedalus Information Systems

21

Monte carlo = rely on repeated, random sampling to predict outcome

# Reinforcement Learning

Program learns reward from human feedback then optimizes reward function

- Rewards:
  - Sampled
  - Evaluative
  - Sequential
- Optimized reward function
- Reward must be explicit to avoid being "gamed"

Issues with tasks and goals

- Too complex
- Hard to specific
- Poorly defined

Daedalus Information Systems

22

Inspired by human decision making

Evaluative feedback is based on decision effectiveness and appropriateness of available alternatives

# Transfer Learning (CS)

Reason relationally

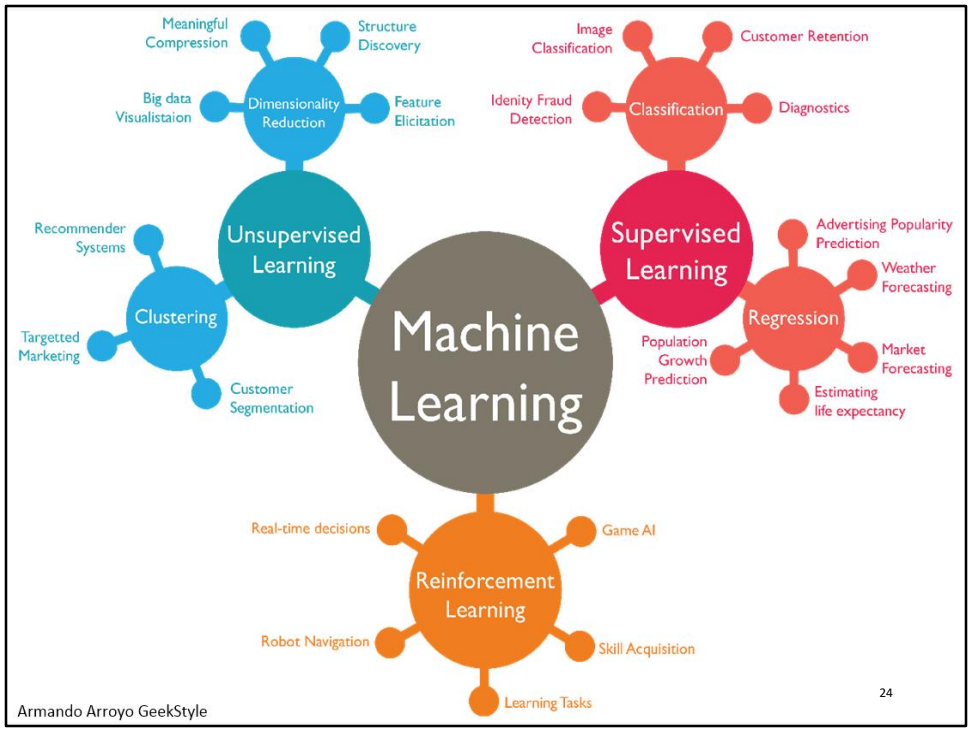
Requires conceptual representation produced by abstract structural knowledge (that is where we humans come in)

Generalizations are transferred to environments that share structures, e.g. mental models

Daedalus Information Systems

23

Imagination key to human ability to plan




https://www.datasciencecentral.com/profiles/blogs/machine-learning-can-we-please-just-agree-what-this-means

# Deep Learning

Collection of trainable math units which collaborate to compute complicated functions

HUGE raw data training set

Results get better with more data, new/better algorithms based on observation and insight

Requirements

- Scalable
- Portable
- Reproducible
- Extensible
- Powerful processing hardware

Daedalus Information Systems

25

"Deep learning, a form of machine learning based on layered representations of variables referred to neural networks, has made speech-understanding practical on our phones and in our kitchens, and its algorithms can be applied widely to an array of applications that rely on pattern recognition" Stanford 100 AI Study
Scalable: sized to meet need
Portable: use across many platforms
Reproducible: by others
Extensible: useful outside of the lab, in a real world environment

"One of the strengths of deep learning is indeed to be able to find feature patterns that humans could probably not predefine. …The developers of facial recognition systems predefine from 50 to 400 measurements and ratios found on the human face (e.g. ratio of distance from eye to nose or nose to lips). In fact facial recognition is completely dependent on these predefined features to work."
https://www.datasciencecentral.com/profiles/blogs/machine-learning-can-we-please-just-agree-what-this-means

Google produces an image of a cat:

- 10 million randomly selected videos with cats along with other subjects
- 16000 processors for parallel processing
- Layering: first layer learns primitive features by finding pixel combinations that

occur more often then should by chance then feeds to next layer on recognized features learned.

- Rinse and repeat.

## The Learning Data Set

If you are not paying for something, you are the product.

Daedalus Information Systems

## What Constitutes a User Profile

Information types

- Demographic
- Interests (short & long-term)
- Preferences

Profiles are dynamic and iterate over time

Represented as

- Set of weighted keyword
- Weighted concepts
- Semantic network

Daedalus Information Systems

27

User profile phases

1. Gather raw information
2. Construct profile from user data
3. Allow application to exploit profile to construct personal results

Keywords profiles represent areas of interest

- Extracted from documents or directly provided by user, weights are numerical representation of user interest
- Polysemy is a big problem for KW profiles

Filtering system

- Network of concepts – unlinked nodes with each node representing a discrete concept
- Used by alta vista (used header that represented user personal data, set of stereotypes (prototypical user comprised of a set of interests represented by a frame of slots
- Each “slot” (made up of domain, topic & weight (domain =area of interest, topic = specific term used to identify area of interest, weight = degree of interest) that makes up frame weighted for relevance

# User Metrics Training Data

Frequency of access
Click-through (selection from results set)
Time on site
Pages per session
Bounce Rate
Conversion  (fulfilled information need)
Profile data

Daedalus
Information Systems

## Implicit Collection

Implicit (max precision 58%)

- Software agents
- Logins
- Enhanced proxy servers
- Cookies
- Session IDs

Gathered without user awareness from behavior

- Query context inferred
- Profile inferred
- Less accurate
- Requires a lot of data

Daedalus Information Systems

29

Jaime Teevan MS Research
(http://courses.ischool.berkeley.edu/i141/f07/lectures/teevan_personalization.pdf)

Tools used

- Software agents: most reliable as more control over install and application
- Cookies: least invasive
- Login: more pervasive across machines and time
- Proxy Servers: limited to user register of machine with server
- Session IDs: limited to a single session

Advantages: more data, better data (easier for system to consume and rationalize)
Disadvantage: user has no control over what is collected

# Explicit Collection

Explicit (max precision 63%)

- HTML forms
- Explicit user feedback interaction (early Google personalization with More Like This)

Provided by user with knowledge

More accurate as user shares more about query intent and interests

Daedalus Information Systems

30

Advantage: User has more control over personal and private information

Disadvantage: compliance, users have a hard time expressing interests, burdensome on user to fill out forms, false info from user

## Google on Privacy

"There was a small trade off on privacy but they're going to get dramatically better search results. That was something that made sense to us over time."

Melissa Mayer
VP User Experience
Google

Daedalus Information Systems

31

www.google.com/history
Web history tied to the Google toolbar (first launched in 2000) and the ability to track what user looked at across the Web

## What Google Collects

| Implicit | Explicit |
| --- | --- |
| Use information | Location information |
| Device information | Profile information |
| Log information | |
| Unique application information | |
| Local storage | |
| Cookie data | |

Daedalus Information Systems 32

Google Privacy Policy http://www.google.com/policies/privacy/
shared across services

- Profile information: Information you give us. For example, many of our services require you to sign up for a Google Account. When you do, we'll ask for personal information, like your name, email address, telephone number or credit card. If you want to take full advantage of the sharing features we offer, we might also ask you to create a publicly visible Google Profile, which may include your name and photo.
- Use information: Information we get from your use of our services. We may collect information about the services that you use and how you use them, like when you visit a website that uses our advertising services or you view and interact with our ads and content. This information includes:
- Device information: We may collect device-specific information (such as your hardware model, operating system version, unique device identifiers, and mobile network information including phone number). Google may associate your device identifiers or phone number with your Google Account.
- Log information "When you use our services or view content provided by Google, we may automatically collect and store certain information in server logs. This may include:
    - details of how you used our service, such as your search queries.
    - telephony log information like your phone number, calling-party number, forwarding numbers, time and date of calls, duration of calls, SMS routing information and types of calls.
    - Internet protocol address.
    - device event information such as crashes, system activity, hardware settings, browser type, browser language, the date and time of your request and referral URL.
    - cookies that may uniquely identify your browser or your Google Account.
- Location information: When you use a location-enabled Google service, we may collect and process information about your actual location, like GPS signals sent by a mobile device. We may also use various technologies to determine location, such as sensor data from your device that may, for example, provide information on nearby Wi-Fi access points and cell towers.
- Unique application numbers" Certain services include a unique application number. This number and information about your installation (for example, the operating system type and application version number) may be sent to Google when you install or uninstall that service or when that service periodically contacts our servers, such as for automatic updates.
- Local storage: We may collect and store information (including personal information) locally on your device using mechanisms such as browser web storage (including HTML 5) and application data caches.
- Cookies and anonymous identifiers: We use various technologies to collect and store information when you visit a Google service, and this may include sending one or more cookies or anonymous identifiers to your device. We also use cookies and anonymous identifiers when you interact with services we offer to our partners, such as advertising services or Google features that may appear on other sites.

## Privacy Paradox

Privacy risk is weighed against value of object, interaction, end result

- Research assumes user calculates an internalized value
- Basis for choice to reveal personal identification information (PII)

Value is determined by the smoothness of the interaction (Groupon, Amazon Local)

- Value proposition overrides security/privacy concerns

Higher level of user control over PII reduces the perception of risk

Daedalus Information Systems

33

Personalization Privacy Paradox: An exploratory study of decision making process for Location-aware marketing: Xu, Luo, Carroll, et.al.
Study focused on location-aware marketing (LAM) – targeting ads, groupons based on awareness of user location, preferences, etc.

Users share private information in exchange for some THING of perceived value and based on assumptions
Agency will deliver (paper, goods, etc.)
They will not share the information indiscriminately
Will protect the data

Users assume a social contract on the part of the agency that they will be responsible

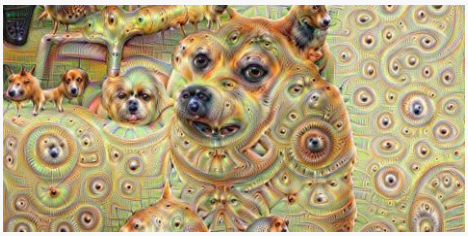
The ease of usability influences the willingness to proceed – Obama campaign online voter registration 2008 – long form split into small, digestible chunks
Previous privacy risk is more influential in covert model (e.g. tracking without user awareness)

# ARTIFICIAL INTELLIGENCE

Not Machine Learning

Daedalus Information Systems

# Norvig & Russell (2004)



Systems that think like humans (neural)

Systems that act like humans (Turing)

System that think rationally (logic solvers)

Systems that act rationally (perception, NLP, Planning, Navigation)

Daedalus Information Systems

35

Artificial Intelligence: A Modern approach

Photo courtesy of https://www.slashgear.com/google-magenta-research-project-will-create-an-ai-artist-23441221/

# Two Schools of AI

Symbol Processing

Artificial Neural Networks

## Symbolic AI

Intelligence = symbol manipulation
Assumes: all intelligent processes are forms of information processing

Intelligent processes = perceiving, reasoning, calculating, language use

Built upon 3 Characteristics of Plato's rationalism

- Psychological
- Epistemological
- Ontological
- [Dreyfus added Biological]

Daedalus Information Systems

37

GOFAI = Fixed and formal rules

Language is symbolic: eg a dog does not look like the word that represents it

Computer processes symbolic representations (1s/0s) according to formal rules (program)

3 characteristics of Plato's rationalism:
Psychological assumption that human intelligence is symbol-manipulation according to formal rules
Epistemological assumption that knowledge is formalized and can be expressed in a context-independent, formal rules or definitions
Ontological that reality has a formalized structure built on objective, determinant elements each of which exists independent of the other .

Dreyfus added the Biological assumption, rules and symbols implemented by the human brain in the same way as by a machine

GOFAI = good old fashioned AI – meat and potatoes AI – train the computer without the need for understanding

## Artificial Neural Networks

Connectionism

Neural networks are made up of input layer, interstitial layer and output layer

Knowledge comes from the connections not symbol interpretation

Past experience used to form intelligence in current state

Pattern recognition, categorization, behavior coordination

Daedalus Information Systems

38

Re-emerged in 1980's
Layers of data – decisions inform up the line (backpropagation)
Autonomy: without human supervision
Automate: replace human effort

Intelligent processing modeled on structure and operation of human brain instead of digital computer – neurons and synapses, receptors and reactors
Neurons as processors with input/output functions
Intelligence is a product of the neuron connections

The ANNs of the 1980s could never conceive of the vast amount of personal and behavioral data used in today's neural networks (deep mind, Watson). Examples: IoT (intelligent machines), Watson (expert systems)

Cannot generalize as humans do, cannot perform functions that require "common sense" (must be programmed)
Heideggerian AI: intelligence is situated in the world and does not require rules.
Terry Winograd (Stanford): design of computers must include consideration that computers must function in a human world and communicate with human users and not impose their own rationalistic logic on surroundings.

# What AI Best Suited To

Search
Learning Systems
Pattern Recognition
Planning
Induction

Daedalus Information Systems

39

Marvin Minsky MIT
Search: search engines
Learning Systems:
Pattern Recognition: fraud detection
Planning: GPS
Induction: IBM Watson

# Search

Requires additional structure

Near to | Close to expansion

Solve for one, Solve for many

Personalization

Daedalus Information Systems

# Pattern Recognition

Ability for computer to act intelligently based on input data with a lot of variability

- Decision Trees
- Nearest neighbor classification
- Neural Networks

Classification

Ideal replaced by practical

Daedalus Information Systems

41

Decision trees: run through series of questions where answer determines outcome

Nearest neighbor: find in training data and use mot similar to predict the unsorted data

Neural networks: based on biochemistry, electric and chemical signals

- some connections dedicated to send, others to receive
- neurons are either idle or firing
- stretch of incoming signals determines the neuron firing
- 2 types of inputs: excitatory (adds up to total) and inhibitory (subtracted from total)
- each neuron assigned a threshold
- signal here is data related to a pre-assigned condition

Explicit teaching based on user data
Learning from example based extracted characteristics from training set of documents

## Planning & Problem-Solving

Large assembly of interrelated sub-problems

Given a start state and desired outcome state

Choose appropriate sub-problems for solving selected problem

Success is most efficient set of actions to achieve desired outcome

Daedalus Information Systems

42

AKA Goal Seeking or Problem Solving
Intelligent systems that decide for themselves
Action and resource management

Given description of start state, a goal state and a sequence of actions. Outcome is to find the most efficient set of actions to achieve the goal
Transportation, scheduling
Interactive decision making: military planning,

# Learning Systems

Use past behavior to predict future action using human planned heuristic methods

A reinforced learning model that leads to a secondary reinforcement model that is more autonomous

- Reinforcement is reward
- Extinction is unlearning

Grade on curve of computer's acquired capability

Daedalus Information Systems

43

Generalized past experiences
Success is reinforced decision models
•Can have secondary reinforcement models (more autonomous)
Reward for partial goals (local reinforcements)
Grade on curve of computers acquired capacity
Reinforcement = reward
Unlearning = extinction

# AI Ethics & Safety

Daedalus Information Systems


Nobody expects
the Spanish Inquisition

## Learning is one Thing…Thinking Another

"In designing software and microprocessors, I have never had the feeling that I was designing an intelligent machine. The software and hardware is so fragile and the capabilities of the machine to "think" so clearly absent that even as a possibility, this has always seemed very far in the future…*My person experience suggest we tend to over estimate our design abilities."*

Daedalus Information Systems

46

Bill Joy, cofounder Sun Microsystems, creator Java and Jini

# Sometimes They Do the Wrong Thing

One of Pieter Bruegel the Elder's 1563 oil paintings of the Towel of Babel

Museum Boijmans Van Beuningen

## An Artificial Intelligence Developed Its Own Non-Human Language

When Facebook designed chatbots to negotiate with one another, the bots made up their own way of communicating.

ADRIENNE LAFRANCE | JUN 15, 2017 | TECHNOLOGY

Daedalus Information Systems

## Algorithmic Bias

Technology inherits ideas and values of the group that develops it

Emotional capitalism: feeling can be managed rationally and governed by logic

Emotional socialism: suffering is unavoidable and should be tolerated

Algorithm development rests on emotional capitalism

Accept decisions from an automated system as agnostic

No universal governance

Daedalus Information Systems

48

Quantified Heart
Polina Aronson
https://aeon.co/essays/can-emotion-regulating-tech-translate-across-cultures

"Algorithms are opinions embedded in code."
Cathy O'Neill, Weapons of Math Destruction (2016)

Ben Schneiderman, winner ACM Turning Award, calls for a national algorithm safety board to monitor and assess safety of algorithms as they access social systems

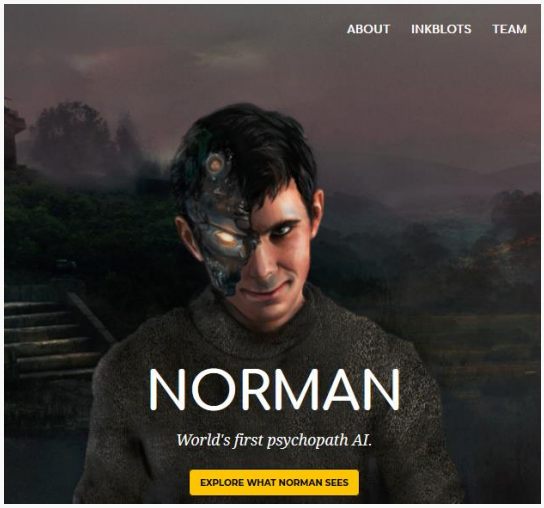
# Sometimes They Learn the Wrong Things

# Sometimes They Build the Wrong Things?

Built as a proof of concept for AI gone wrong with biased data

MIT AI Lab

Dataset was a sub-reddit dedicated to document the "disturbing reality of death."




Daedalus Information Systems

50

http://norman-ai.mit.edu/ -

"Norman suffered from extended exposure to the darkest corners of Reddit, and represents a case study on the dangers of Artificial Intelligence gone wrong when biased data is used in machine learning algorithms. "

Also produced Shelley (http://shelley.ai/), AI assisted horror stories, and Deep Empathy (https://deepempathy.mit.edu/) that produces images of what US cities would look like after conflict similar to that experienced in Syria

# Adversarial AI


Used in for negative outcomes

- Autonomous weapons
- Biased facial recognition

Used for malicious purposes

- Fake news
- Denial of attack

Daedalus Information Systems

51

Adversarial setting example: Russian hacking of US election
Malicious means: Fall 2016 IoT hack that took down part of the internet

# AI Risks

Mis-specified Objectives

Negative Side Effects that extend to wider application

Hacking: rewards, devices

Bad extrapolation of the real world

Poor training data

Privacy

Fairness

Abuse

Transparency

Daedalus Information Systems

52

Google Report on AI Safety
Privacy: right to be forgotten
Fairness: digital divide
Security: IoT takedown of internet, GM self-driving car
Transparency: common understanding of complex engineering

# Governance Issues

Explanation *(transparency)*

- Core components
- Local Explanation: explain for specific decision, not system as a whole
- Counterfactual Faithfulness: expect the explanation to be causal and can be provided without providing contents of the system
- Provide in situations where a person would be required to do so

Regulation

- Regulators don't understand what they are regulating
- Risk of stifling innovation

Applications *(consistency)*

- Impact beyond decision-maker
- Know if AI behaving erroneously

Daedalus Information Systems

53

Accountability of AI Under the Law: Doshi-Velez, Kortz, et.al. : Harvard; 2017

Explanation is different than transparency – should not require knowledge of the "flow of bits through AI"
No hiding behind the technology

Nobody knows anything..... Not one person in the entire motion picture field knows for a certainty what's going to work. Every time out it's a guess and, if you're lucky, an educated one.

William Goldman

54

Ask why.
Why is this important?
Why should you care?
Why will it be effective?

# WHERE WE COME IN

IA to the rescue

Daedalus
Information Systems

55

# AI Risk Mitigations

Define impact regulator

- Future state
- Substitutes lower impact null actions

Train impact regulator

- Over many tasks
- Separate training parameters for task side effects

Penalize influence

- Use information-theoretic measures to capture agent's potential for information
- Penalize empowerment

Provide scalable oversight with multi-agent approach

Daedalus Information Systems

56

Multi-agent approach = human and agent working together

Reward Hacking: adversarial reward function, careful programming to avoid adversarial blindness

Scalable oversight: distant supervision, hierarchical reinforcement learning

# AI Risk Mitigations 2

Use Objective functions to capture designer informal intent

- No partially observed goals
- Concrete, not abstract rewards
- Deep correlation between tasks and functions

Feedback loops

- Model look ahead
- Reward capping
- Counter example resistance – combination of rewards

Daedalus Information Systems

57

Correlation between tasks and rewards: do not base cleaning robot reward on amount of cleaning supplies used

# AI Risk Mitigations 3

Safe exploration

- Risk sensitive performance criteria
- Use demonstration
- Simulated exploration

Well defined models

- Train on multiple distributions
- Program for out-of-distribution situations

Daedalus Information Systems

58

Simulated exploration: bounded exploration, trusted policy oversight, human oversight

# Information Architecture and AI

Problem definition and structure

Connections

Proto-typicality (mental models)

Visual complexity (rely on text more than images)

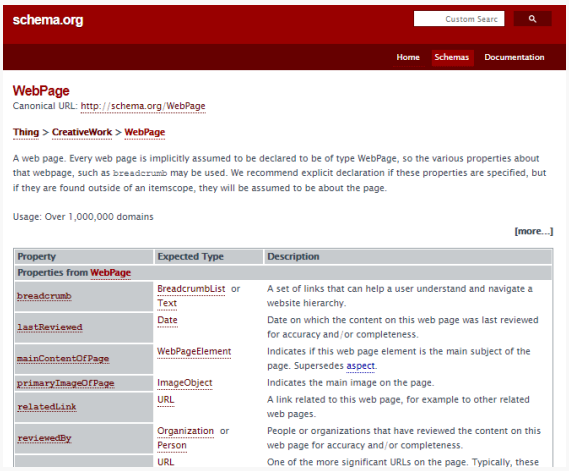

59

Legacy newspaper structure of “the fold.”

Proto-typicality: user mental models

Visual complexity: ratio of images to text favors text

# Structured Data



schema.org

Custom Searc

Home Schemas Documentation

## WebPage

Canonical URL: http://schema.org/WebPage

Thing > CreativeWork > WebPage

A web page. Every web page is implicitly assumed to be declared to be of type WebPage, so the various properties about that webpage, such as breadcrumb may be used. We recommend explicit declaration if these properties are specified, but if they are found outside of an itemscope, they will be assumed to be about the page.

Usage: Over 1,000,000 domains

[more...]

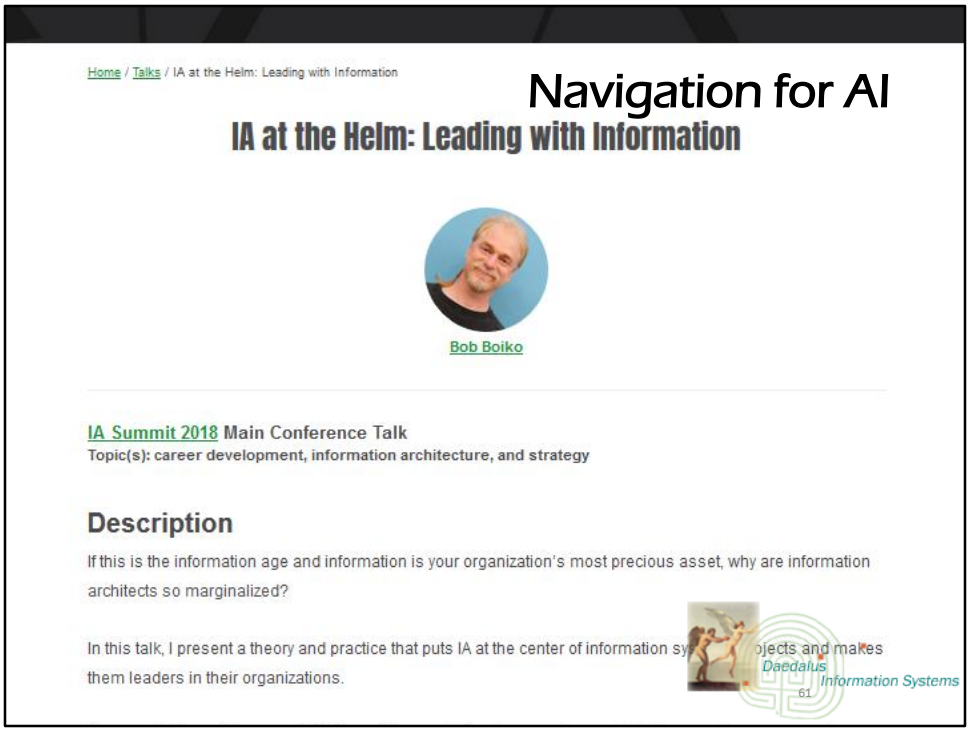
| Property | Expected Type | Description |
|---|---|---|
| **Properties from WebPage** | | |
| breadcrumb | BreadcrumbList or Text | A set of links that can help a user understand and navigate a website hierarchy. |
| lastReviewed | Date | Date on which the content on this web page was last reviewed for accuracy and/or completeness. |
| mainContentOfPage | WebPageElement | Indicates if this web page element is the main subject of the page. Supersedes aspect. |
| primaryImageOfPage | ImageObject | Indicates the main image on the page. |
| relatedLink | URL | A link related to this web page, for example to other related web pages. |
| reviewedBy | Organization or Person | People or organizations that have reviewed the content on this web page for accuracy and/or completeness. |
| | URL | One of the more significant URLs on the page. Typically, these |

Name the components on the page for the machine user

Daedalus Information Systems

60

https://schema.org/WebPage




6-part series called **Information Systems from the Info Out**

# Map Semantic Connections

Semantic technology requires everything to be associated to understand user activity

- Control layer
- Mapping (semantic) layer
- Device layer

Semantic analysis model

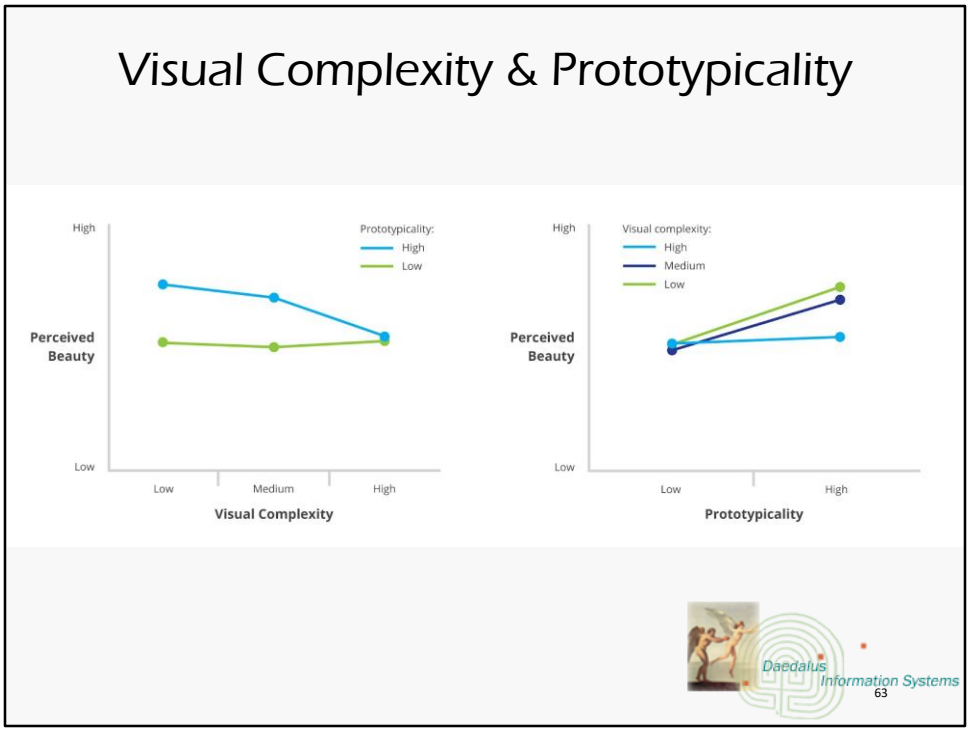
- Semantic layering
- Semantic mapping (Boiko IAS 2018)
- Semantic machine heterogeneity

Association between user behavior patters (customer journey map)

Daedalus Information Systems

62

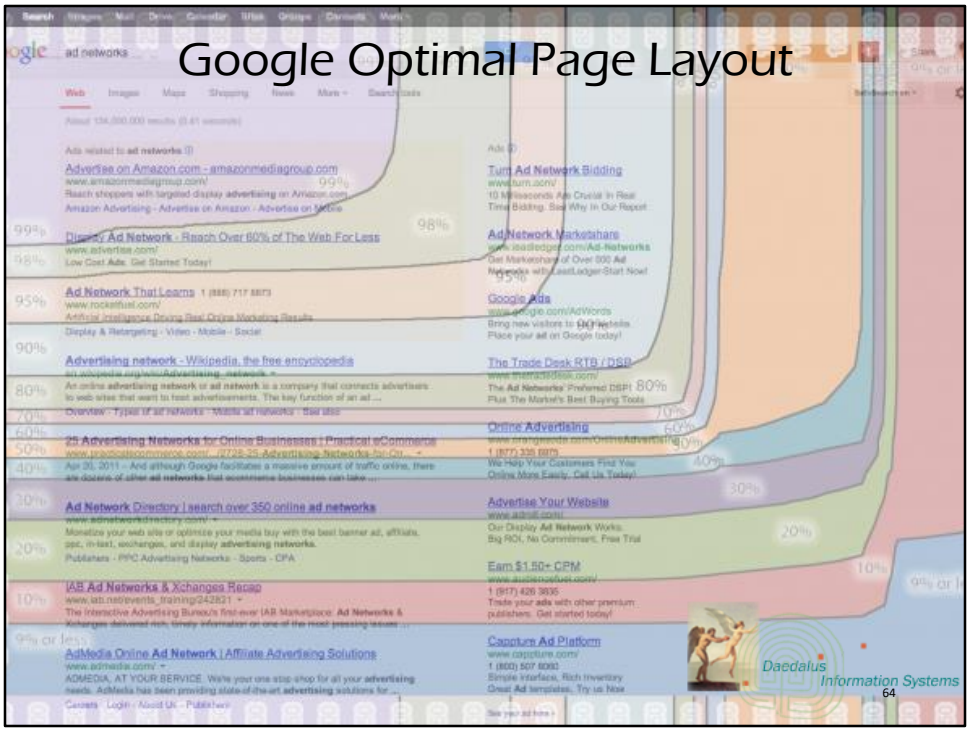
Useful tools here
User models (personas): characteristic preferences
Knowledge models (journey maps): information behavior




VISUAL COMPLEXITY & PROTOTYPICALITY
The results show that both visual complexity and proto-typicality play crucial roles in the process of forming an aesthetic judgment. It happens within incredibly short timeframes between 17 and 50 milliseconds. By comparison, the average blink of an eye takes 100 to 400 milliseconds.

In other words, users strongly prefer website designs that look both *simple* (low complexity) and *familiar* (high prototypicality). That means if you're designing a website, you'll want to consider both factors. Designs that contradict what users typically expect of a website may hurt users' first impression and damage their expectations.
August 2012

Resource: http://googleresearch.blogspot.com/2012/08/users-love-simple-and-familiar-designs.html




From Patent: Techniques for approximating the visual layout of a web page and determining the porting of the page containing significant content.

“As we’ve mentioned previously, we’ve heard complaints from users that if they click on a result and it’s difficult to find the actual content, they aren’t happy with the experience. Rather than scrolling down the page past a slew of ads, users want to see content right away. So sites that don’t have much content “above-the-fold” can be affected by this change.”
http://googlewebmastercentral.blogspot.com/2012/01/page-layout-algorithm-improvement.html

Resources
http://www.seobythesea.com/2011/12/10-most-important-seo-patents-part-3-classifying-web-blocks-with-linguistic-features/
http://www.seobythesea.com/2008/03/the-importance-of-page-layout-in-seo/

# Form IA and AI Strategies

Customer Empathy Framework

- Define the problem
- Formulate the solution
- Map the environment (customer journey)

Tools

- Personas (use cases)
- Problem statements
- Environment description (include systems and processes)
- Success benchmark success (quantitative, qualitative)

Daedalus Information Systems

65

https://www.linkedin.com/pulse/design-thinking-data-science-george-roumeliotis
http://www.intuitlabs.com/page/2/?s=design+for+delight

# Use a Different Pattern Library

Visitor search patterns: Use online tools to uncover customer intent

Visitor behavior patterns: website analytics

Visitor conversion patterns: content to address all stages of conversion funnel

Tools

- Search suggest scrapers
- SEO|Content Marketing software
- Webmaster and website analytics accounts

# Transform Keywords Into Intelligence

Keywords are user queries

Queries represent user information needs and satisfaction threshold

Keywords become intelligence

- Competitive: who is doing better
- Visibility: how do the search engines see my content
- Customer: how do targeted customers look for my products and Services

Tools

- Search suggest scrapers
- Google Trends
- SEO Software (BrightEdge, SEMrush)

Daedalus Information Systems

# Create Meaningful Structures

Site Structure

- Machine readable text
- Related content model
- Schema markup

Internal linking to reinforce context relationships and discovery




Legacy newspaper structure of “the fold.”

Proto-typicality: user mental models

Visual complexity: ratio of images to text favors text

## Create and Curate Content

Entities Rule

Newspaper model

Opening paragraphs most important for subject determination

Relational content model

Daedalus Information Systems

69

"things not strings" Amit Singhal

Deep, rich content is rewarded with higher visibility

More content = Authority = aboutness

People will scroll - If they don't scroll, they will print it out

Visible text on a page is what counts

Spiders cannot "see" = cannot read text images

Relational content model – tie contextually relevant content together for visitors.
Don't make them use search engines.
Search engines are:
Semantic (LSI)
Judgmental
Evaluate content based on non-content criteria (bounce rate, click through, conversion)

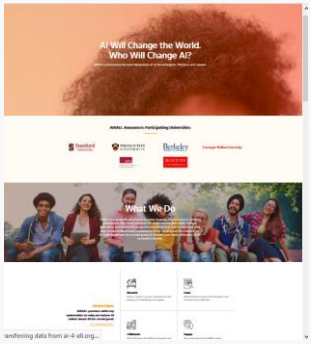
# AI NOW Initiative (2018)

Kate Crawford (Microsoft) and Meredith Whittaker(Google)

Founded to deal with issues of AI diversity and inclusion

Conduct empirical studies focused on

- Bias and inclusions
- Labor and automation
- Infrastructure and Safety
- Basic rights and liberties




Daedalus Information Systems

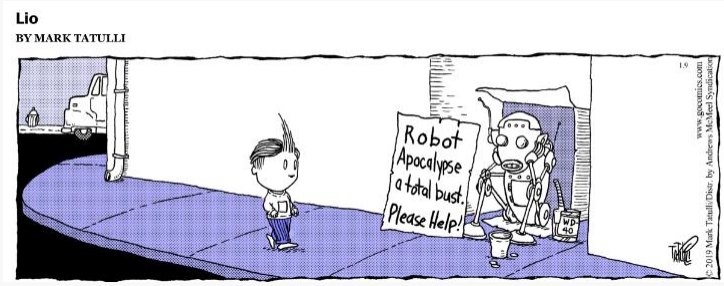
70

http://ai-4-all.org/


Photo courtesy of https://www.chemistryworld.com/news/what-makes-a-snowflake-special/3008386.article

# Thank You

Embrace, engage, define, direct




Marianne Sweeny
Principal
Daedalus Information Systems
sweeny48@uw.edu
@msweeny

Daedalus Information Systems

72

We can circumvent the AI apocalypse.

# APPENDIX

A friendly drop to help you after...

Daedalus
Information Systems

# Suggested Reading

- *Algorithms to Live By*; Brian Christian, Tom Griffiths
- Super Intelligence: Paths, Dangers, Strategies; Nick Bostrum
- The Tides of Mind: Uncovering the Spectrum of Consciousness; David Gelernter
- The Knowing Project; Michael Lewis

Daedalus Information Systems

# Twitter Resources for User-Centered AI

Rob Wortham @RobWortham
Frank Pasquale @FrankPasquale
John C. Havens @johnchavens
Joanna Bryson @j2breve, @j2blather
Carol Smith @carologic
Sentiment/Emotion/AI @SentimentSymp
Elizabeth Churchill @xeeliz
Adam Coates @adampaulcoates
Richard @RichardSocher
Yann LeCun @ylecun
Andrew Ng @AndrewYNg
Eric Horvitz @erichorvitz
Oren Etzioni @etzioni
Jeff Dalton @JeffD
Kevin Slavin @slavin_fpo
Giles Colborne @gilescolborne
Rob McCargow @robmccargow

Dorian Taylor @doriantaylor
Dave Snowden @snowded
Jana Eggers @jeggers
Dawn Anderson @dawnieando
Kirk Borne @KirkDBorne
Colin Eagan @ColinEags
Data Science Central @DataScienceCtrl
Right Relevance @rightrelevance
Machine Learning @ML_toparticles
Tim Caynes @timcaynes
Brenda Laurel @blaurel
Ian Soboroff @ian_soboroff
Phillip Hunter @designoutloud
Paul Dourish @dourish
Kate Crawford @katecrawford
Me @msweey

Daedalus Information Systems