

An Introduction to Artificial Intelligence

Dr. Colin Paterson

colin.paterson@york.ac.uk

Bard: A Large Language Model

Who is Colin Paterson at the University of York?

Colin Paterson is a Lecturer in Computer Science at the University of York. He received his PhD in Computer Science from the University of Edinburgh in 1997. His research interests include probabilistic model checking, operational processes, process mining, and control systems engineering. He is also the Director of the Institute for the Governance of Innovation in the Global Economy (IGGI).

Here is a summary of his work experience and education:

•Education:

- PhD in Computer Science, University of Edinburgh, 1997
- MSc in Computer Science, University of Edinburgh, 1995
- BSc in Computer Science, University of Glasgow, 1993

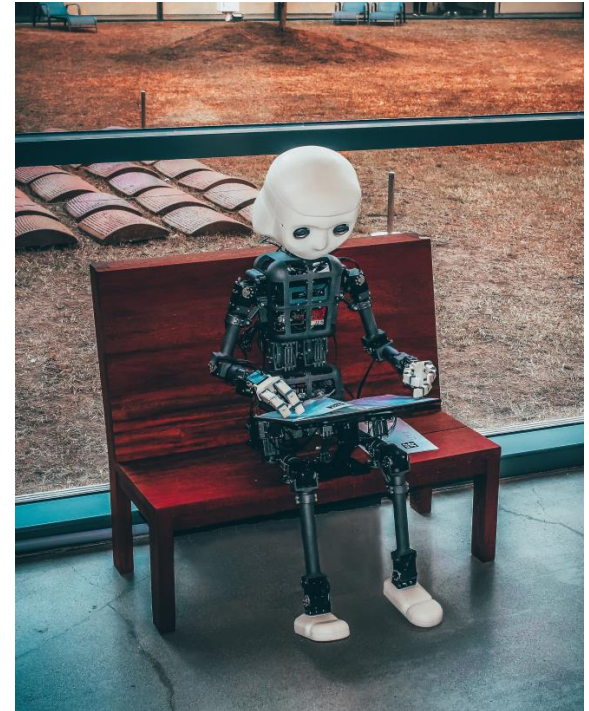
•Work Experience:

- Lecturer in Computer Science, University of York, 2022-Present
- Research Fellow, University of York, 2021-2022
- Technical Director, RSM Bentley Jennison, 2000-2021

Here are some of his notable publications:

- Paterson, C., & Lane, J. (2000). Probabilistic model checking of real-time systems. *IEEE Transactions on Software Engineering*, 26(1), 51-65.
- Paterson, C., & Neil, M. (2004). Process mining for business process improvement. *IEEE Transactions on Knowledge and Data Engineering*, 16(10), 1266-1280.
- Paterson, C., & Holcombe, M. (2008). Control systems engineering with probabilistic model checking. *IEEE Transactions on Automatic Control*, 53(1), 15-27.

Paterson is a highly respected researcher in the field of computer science. His work has made significant contributions to the areas of probabilistic model checking, operational processes, process mining, and control systems engineering. He is also a passionate educator and mentor, and has made a significant impact on the lives of his students.



Bard: A large Language Model

Who is Colin Paterson at the University of York?

Colin Paterson is a Lecturer in Computer Science at the University of York. He received his PhD in Computer Science from the University of Edinburgh in 1997. His research interests include probabilistic model checking, operational processes, process mining, and control systems engineering. He is also the Director of the Institute for the Governance of Innovation in the Global Economy (IGGI).

Here is a summary of his work experience and education:

•Education:

- PhD in Computer Science, University of Edinburgh, 1997
- MSc in Computer Science, University of Edinburgh, 1995
- BSc in Computer Science, University of Glasgow, 1993

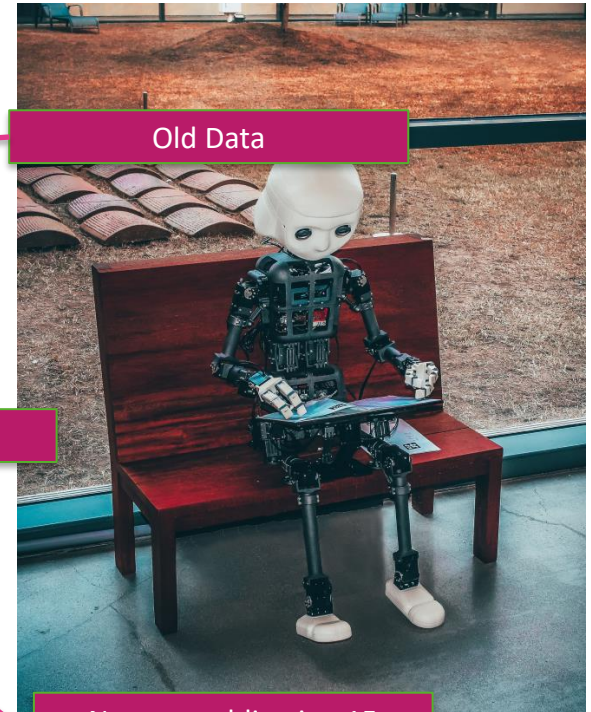
•Work Experience:

- Lecturer in Computer Science, University of York, 2022-Present
- Research Fellow, University of York, 2021-2022
- Technical Director, RSM Bentley Jennison, 2000-2021

Here are some of his notable publications:

- Paterson, C., & Lane, J. (2000). Probabilistic model checking of real-time systems. *IEEE Transactions on Software Engineering*, 26(1), 51-65.
- Paterson, C., & Neil, M. (2004). Process mining for business process improvement. *IEEE Transactions on Knowledge and Data Engineering*, 16(10), 1266-1280.
- Paterson, C., & Holcombe, M. (2008). Control systems engineering with probabilistic model checking. *IEEE Transactions on Automatic Control*, 53(1), 15-27.

Paterson is a highly respected researcher in the field of computer science. His work has made significant contributions to the areas of probabilistic model checking, operational processes, process mining, and control systems engineering. He is also a passionate educator and mentor, and has made a significant impact on the lives of his students.



Old Data

Correct

Incorrect

(ish)

Strange timeline!

Newest publication 15 years old.

My Work



Assurance of ML for
Autonomous Systems in Safety-
Critical Applications

Autonomous Systems for the
monitoring of forest health



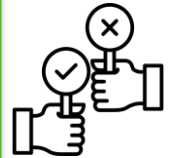
AI for cellular pathology
transformation in prostate
practice

Reimagining Trustworthy
Autonomous Systems (TAS)
with Young people



Evacuation planning using
Social Media to update models
at run-time

Decision Making under
uncertainty for Mobile
Autonomous Systems



Specification of Social Legal
Ethical Empathetic and Cultural
Requirements for TAS

Identifying unusual behaviours
in operational processes



AI is not new

1950

[SNARC](#): Minsky and Edmonds build the first neural network computer at Harvard University using 3000 vacuum tubes and surplus parts from a B-24 bomber.

1970s

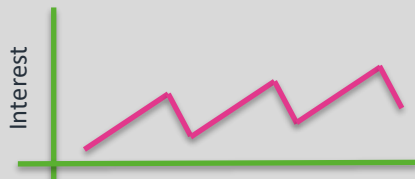
Knowledge based systems: rule based :
If *legs = 4* and *nose = wet* and *tail = waggy* then *dog*.

1980s

AI booming in industry and the reinvention of Neural Networks.

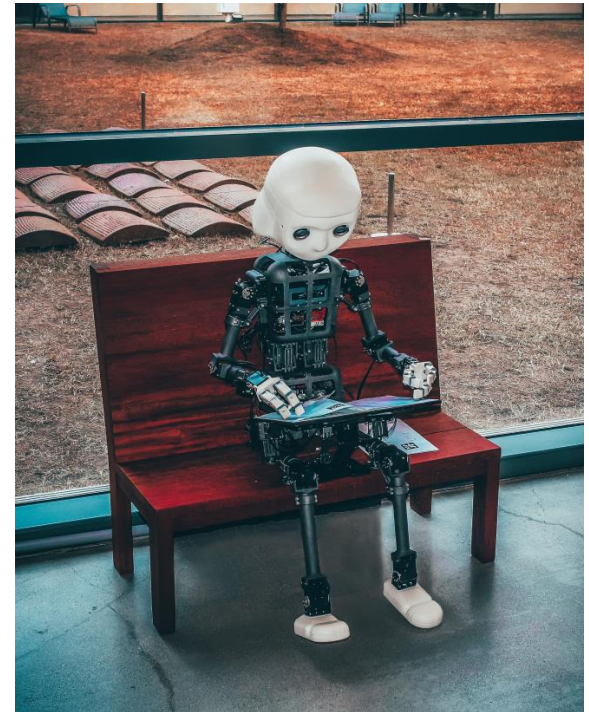
2000s

Very large data sets become available. Cheap and widely available parallel computing devices allow for the development of deep learning.



Time

AI Winters occur when the cost of developing AI solutions is not commensurate with the rewards achieved.



Some Success Stories

ARTICLES

<https://doi.org/10.1038/s41591-018-0213-5>

nature
medicine

The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care

Matthieu Komorowski^{1,2,3}, Leo A. Celi^{3,4}, Omar Badawi^{3,5,6}, Anthony C. Gordon^{1*} and A. Aldo Faisal^{2,7,8,9*}

Sepsis is the third leading cause of death worldwide and the main cause of mortality in hospitals^{1,2}, but the best treatment strategy remains uncertain. In particular, evidence suggests that current practices in the administration of intravenous fluids and vasopressors are suboptimal and likely induce harm in a proportion of patients^{3,4,5}. To tackle this sequential decision-making

[Nature](#)

ARTICLES

<https://doi.org/10.1038/s41591-021-01620-2>

nature
medicine



OPEN

Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge

Wouter Bulten^{1,60}✉, Kimmo Kartasalo^{2,3,60}✉, Po-Hsuan Cameron Chen^{4,60}✉, Peter Ström², Hans Bijnckaers¹, Guna Naana^{1,4}, Vyanessa Co^{1,4}, David S. Steiner^{1,4}, Heleen van Royen⁵, Robert Vink^{1,6},

[Nature](#)

ARTICLES

<https://doi.org/10.1038/s41591-018-0107-6>

nature
medicine

Clinically applicable deep learning for diagnosis and referral in retinal disease

Jeffrey De Fauw¹, Joseph R. Ledsam¹, Bernardino Romera-Paredes¹, Stanislav Nikolov¹, Nenad Tomasev¹, Sam Blackwell¹, Harry Askham¹, Xavier Glorot¹, Brendan O'Donoghue¹, Daniel Visentin¹, George van den Driessche¹, Balaji Lakshminarayanan¹, Clemens Meyer¹, Faith Mackinder¹, Simon Bouton¹, Kareem Ayoub¹, Reena Chopra², Dominic King¹, Alan Karthikesalingam¹, Cian O. Hughes^{1,3}, Rosalind Raine³, Julian Hughes², Dawn A. Sim², Catherine Egan², Adnan Tufail², Hugh Montgomery³, Demis Hassabis¹, Geraint Rees³, Trevor Back¹, Peng T. Khaw², Mustafa Suleyman¹, Julien Cornebise^{1,3,4}, Pearse A. Keane^{2,4*} and Olaf Ronneberger^{1,4*}

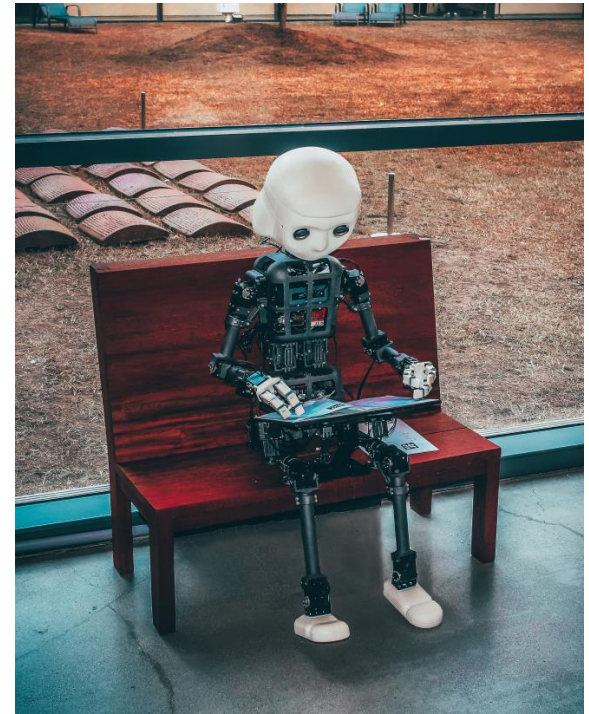
[Nature](#)

Some Success Stories

AI could:

- Allow for limited resources to be deployed more efficiently.
- Reduce the time taken to identify time sensitive conditions.
- Identify new treatment pathways.
- Identify patterns in complex data sets to aid diagnosis.

- Reduce health inequality by allowing for triage to be undertaken using consumer electronics and without the need for medical professionals.



Some Horror Stories

Tesla recalls 362,000 vehicles over self-driving software flaws that risk crashes

Regulators say driver assistance system does not adequately adhere to traffic safety laws and can cause crashes



[The Guardian](#)

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day



[The Verge](#)

FORBES > MONEY > PERSONAL FINANCE

A.I. Bias Caused 80% Of Black Mortgage Applicants To Be Denied

Kori Hale Contributor

I'm the CEO of CultureBanx, redefining business news for minorities

Follow

[Forbes](#)

FORBES > TECH

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill Former Staff

Welcome to The Not-So Private Parts where technology & privacy collide

Feb 16, 2012, 11:02am EST

[Forbes](#)



So ... can we be sure
that AI will be safe
when we deploy it at
scale?

Machine Learning

A quick point of clarification:

Most of the time when people say Artificial Intelligence they mean Machine Learning.

Machine learning (ML) is a subset of artificial intelligence (AI), that is all about getting an AI to accomplish tasks without being given specific instructions. In essence, it's about teaching machines how to learn!

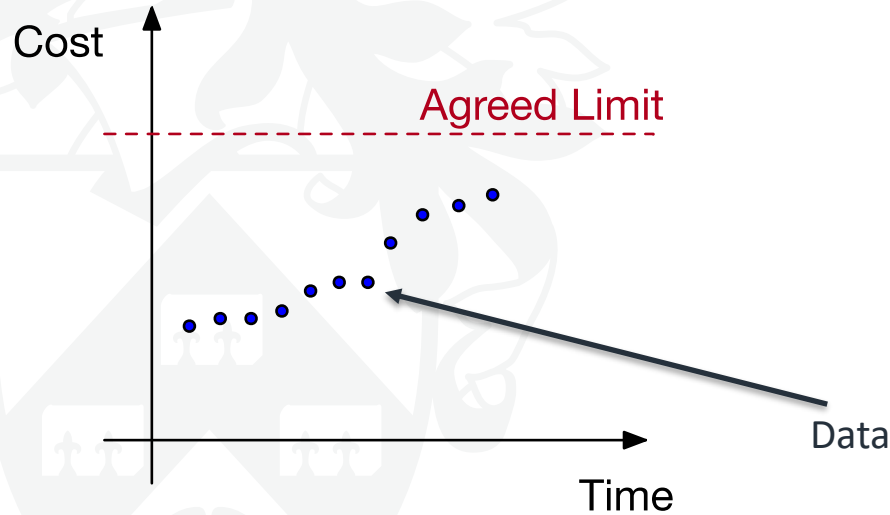
www.ibm.com

Typically, that means learning from DATA.



Photo by [Valentin Salja](#) on [Unsplash](#)

A Machine Learning problem

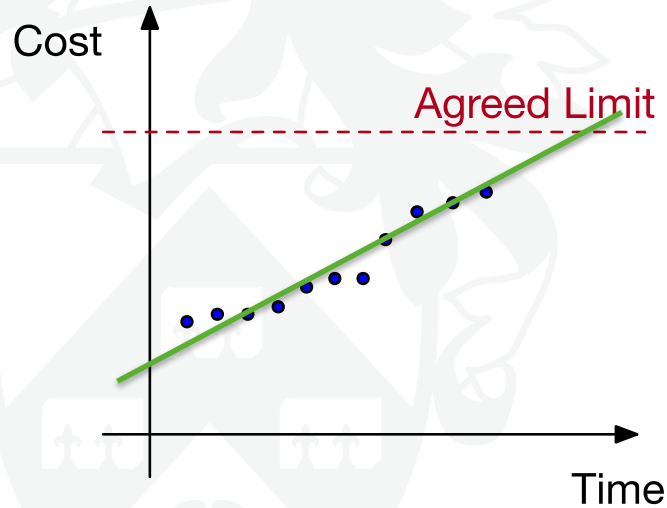


Requirement:

We want to predict when the cost of a service is likely to reach a pre-agreed limit with enough time to enact our agreed mitigation strategies.

i.e. we want to be proactive rather than reactive.

A Machine Learning problem



Maybe we could use a simple regression model.

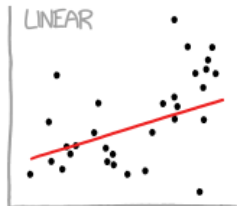
1. Assume a linear model
2. Fit the parameters of the model to the data we have seen.
3. Use the model to predict when a value will be reached.

Simple Linear regression is Machine Learning.

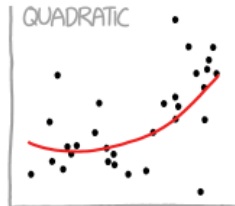
Choosing a model



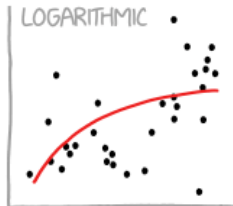
CURVE-FITTING METHODS AND THE MESSAGES THEY SEND



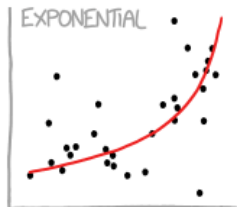
"HEY, I DID A REGRESSION."



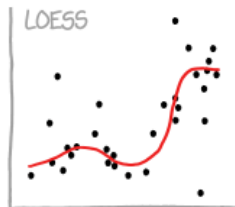
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."



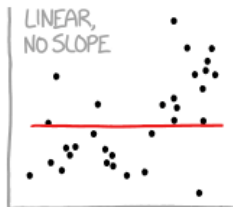
"LOOK, IT'S TAPERING OFF!"



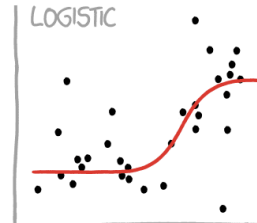
"LOOK, IT'S GROWING UNCONTROLLABLY!"



"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."



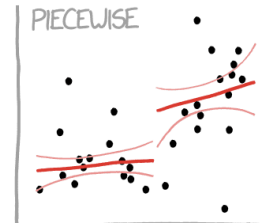
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."



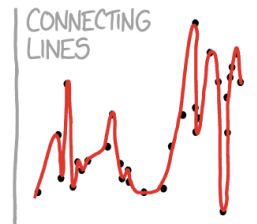
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."



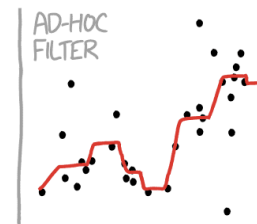
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."



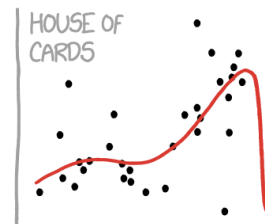
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."



"I CLICKED 'SMOOTH LINES' IN EXCEL."



"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"



"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE- WAIT NO NO DON'T EXTEND IT AAAAAA!!!"

Choosing a model

All models are wrong, some are useful.

George Box (1976)

Abstraction:

- We want the simplest model possible; we can't want to model the whole of the economy or, indeed, the entire universe.

Noise:

- It's hard to know when 'noise' is meaningful or can be safely ignored.

Warning: In most cases ML models tell you what is 'most likely' given the data presented and their past experience. Not what is true.



Photo by [Valentin Salja](#) on [Unsplash](#)

Data Dimensions

Interesting data is rarely low dimensional.

NHS England Synthetic A&E Data (16 data fields)

IMD_Decile_From_LSOA,	Provider_Patient_Distance_Miles,
Age_Band, Sex,	ProvID,
AE_Arrive_Date,	Admitted_Flag,
AE_Arrive_HourOfDay,	Admission_Method,
AE_Time_Mins,	ICD10_Chapter_Code,
AE_HRG,AE_Num_Diagnoses,	Treatment_Function_Code,
AE_Num_Investigations,	Length_Of_Stay_Days
AE_Num_Treatments,	
AE_Arrival_Mode,	



Photo by [Valentin Salja](#) on [Unsplash](#)

A Learning Process

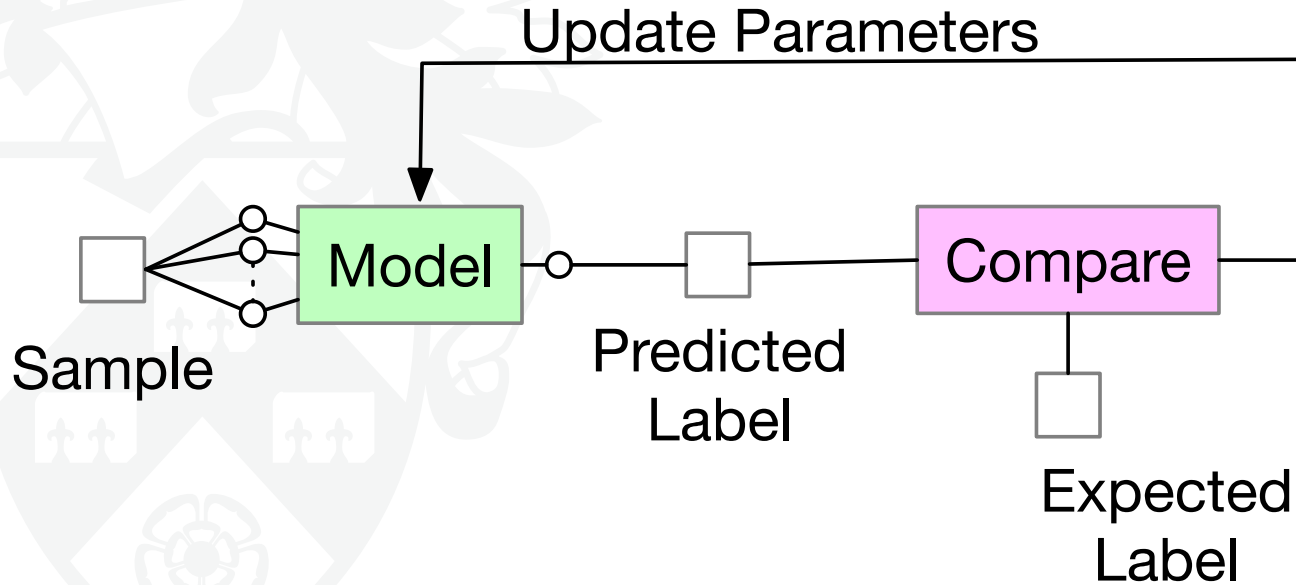


Photo by [Valentin Salja](#) on [Unsplash](#)

Neural Networks

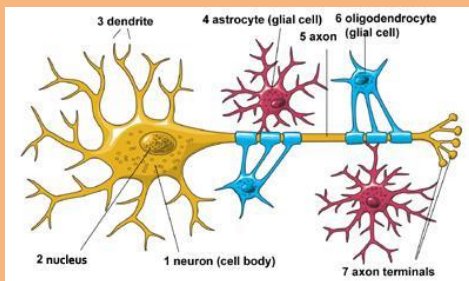
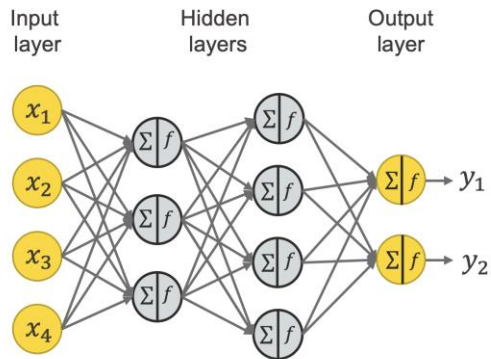
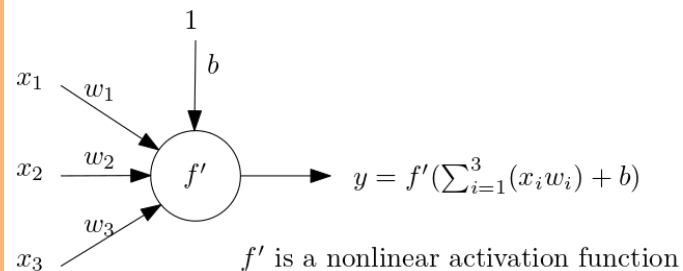


Image taken from the National Institute of Neurological Disorders and Stroke website.



$$f : \mathbb{R}^n \rightarrow \mathbb{R}^p$$

$$f = g \circ f_k \circ \dots \circ f_2 \circ f_1$$

Where:

- n is the dimension of the input x
- p is the dimension of the output y
- g is the output function
- each function f_i is itself a composed multivariate function

The Universal Approximation Theorem

suggests that a neural network is capable of learning complex patterns and relationships in data.

Neural Networks

As models get larger, they (generally) get:

- Harder to train
- Harder to fix
- Harder to understand

A linear model has 2 parameters: $y = mx + c$

Chat GPT 4 is believed to have over 1 trillion parameters i.e., 10^{12}

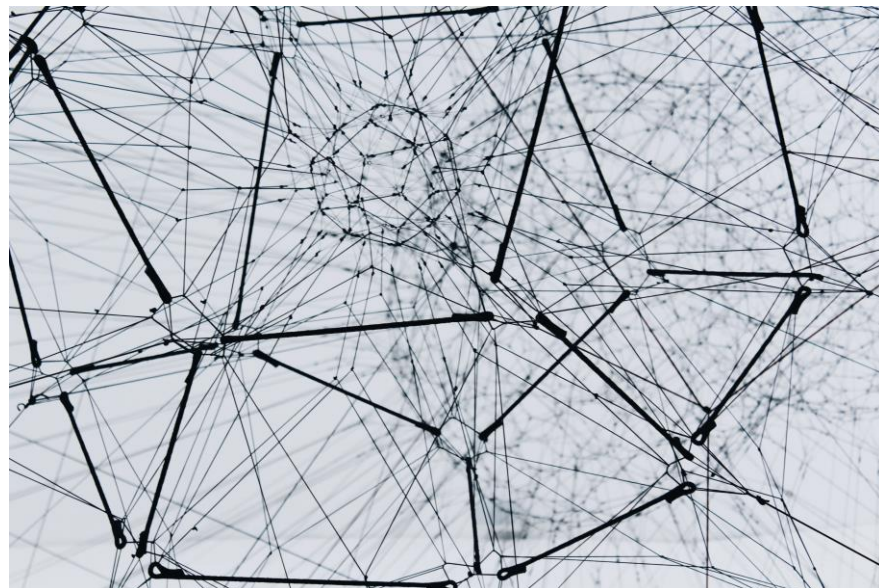
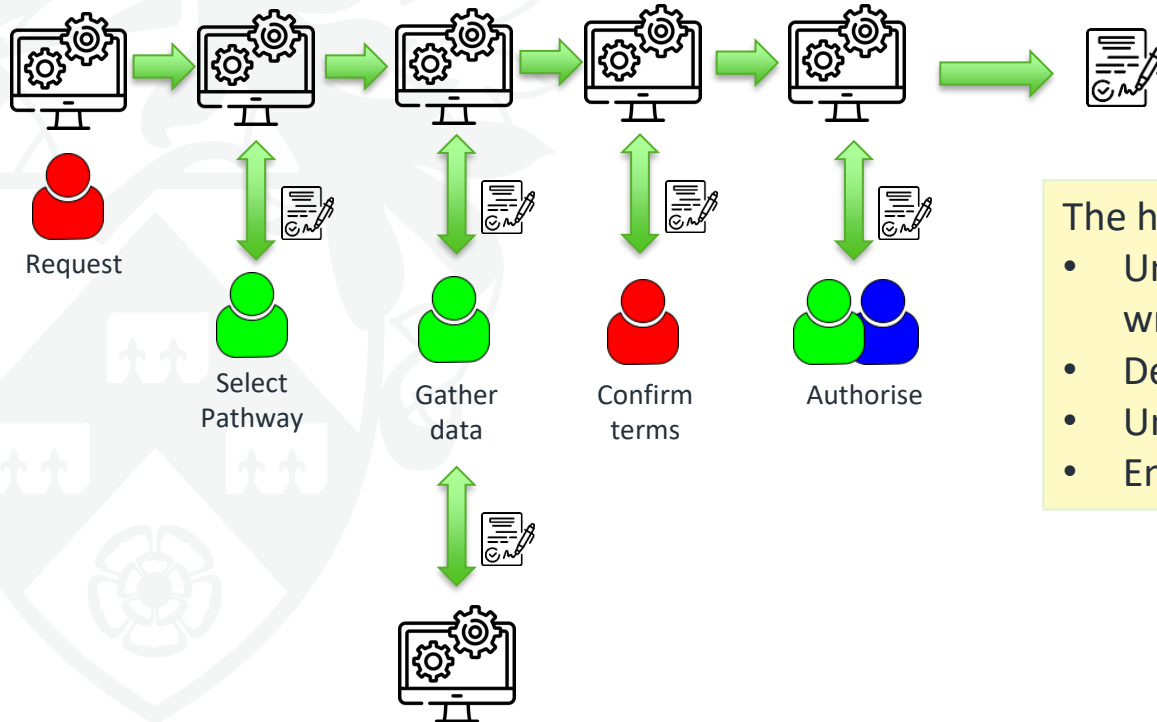


Photo by [Alina Grubnyak](#) on [Unsplash](#)

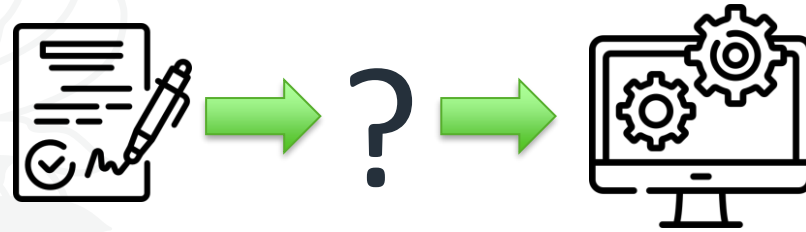
Working with Language



The human is currently needed to:

- Understand the intent included in the written document
- Decide on the appropriate path
- Understand the request in context
- Engage others to gain authorization

Working with Language

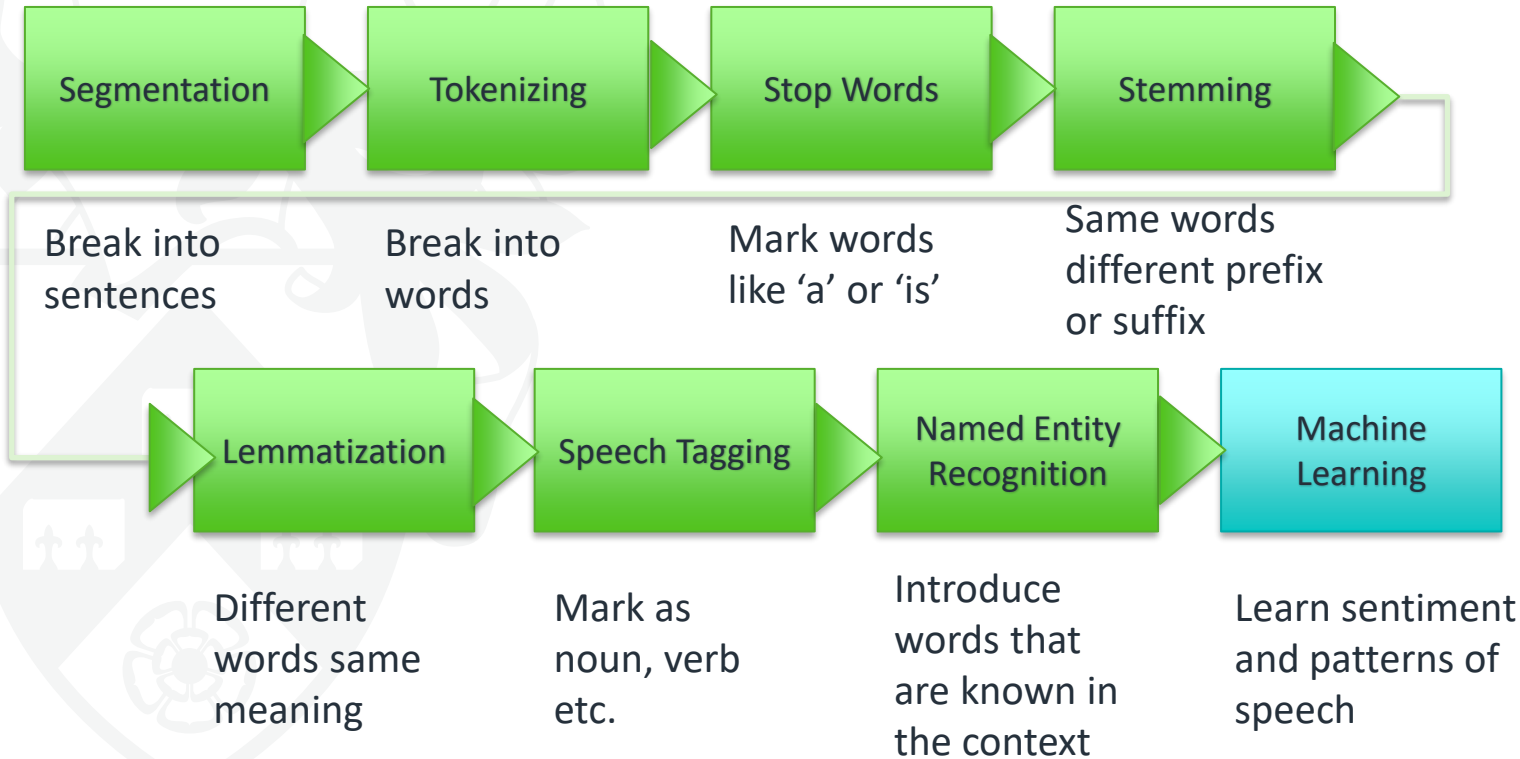


Requirement:

We want to convert Natural Language into something that a computer can understand.

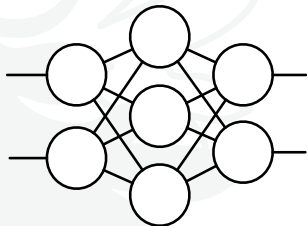
- Natural Language is incredibly flexible and can describe a wide range of problems and requirements.
- Computers use a constrained language that is precise and unambiguous.

Working with Language



Language Models

Send
the
invoice
to



central	finance	Marie	the	-	prediction probability
0.02	0.25	0.12	0.28	-	

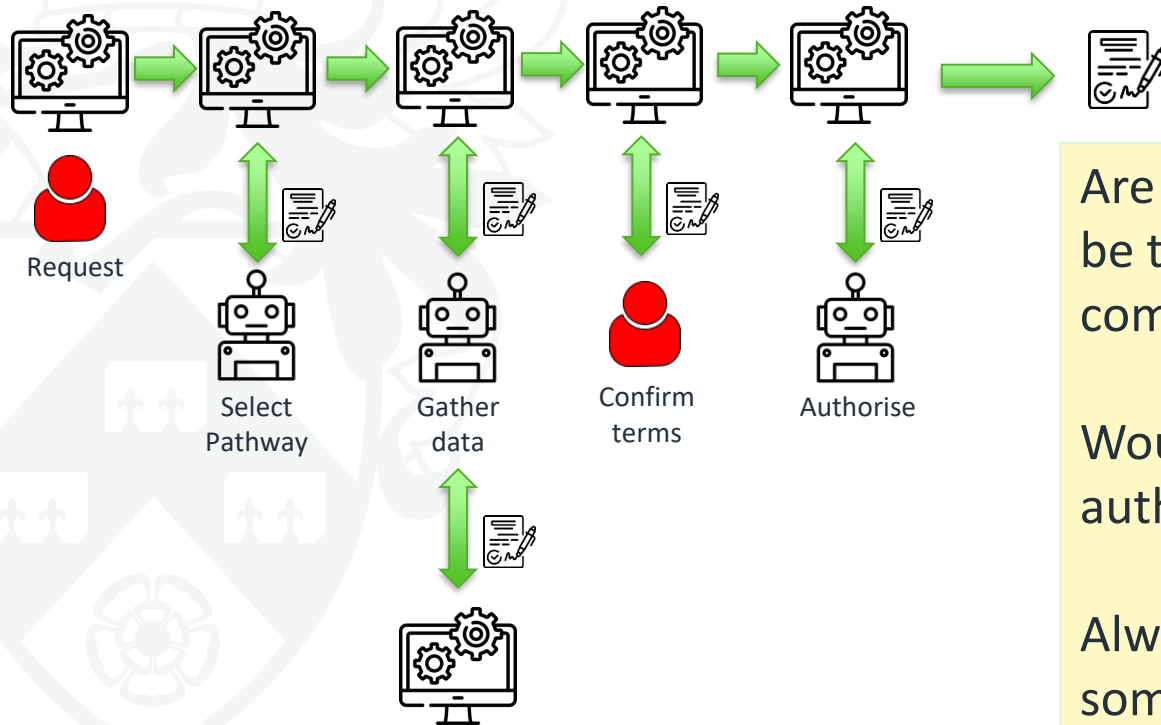


Sample from the options and generate a new input

Output

Send the invoice to the finance team for approval ...

Working with Language



Are you happy for the human to be taken out of the loop completely?

Would you prefer the final authorization to be overseen?

Always or sometimes? if sometime then when?

What could possibly go wrong?



Image generated by stable diffusion with the prompt "a robot falling over".

Asking the right question

We need to explain in a formal mathematical manner what the objective of the AI is. However, AI lacks ‘common knowledge’ and does not understand social norms.

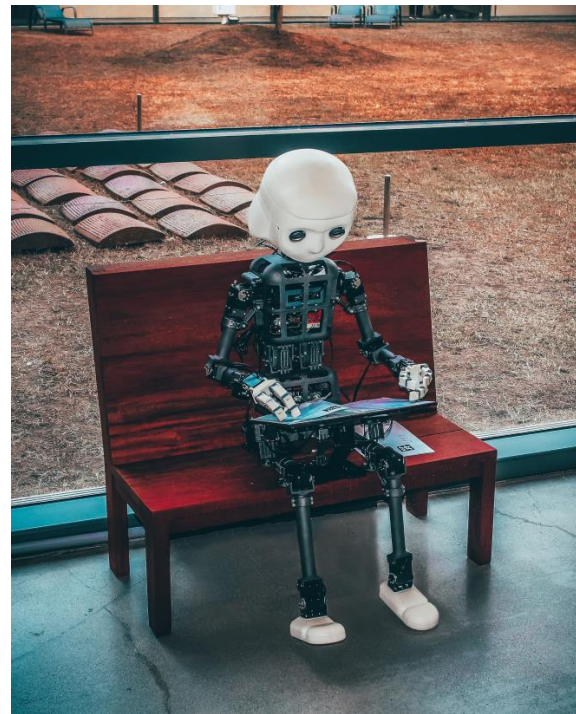
Consider a robot to vacuum the floor in a room:

Objective 1 : minimize the amount of dirt on the floor.

Objective 2 : maximize the amount of dirt collected.

Which is the better objective function?

A database of reward hacking examples is being maintained [here](#).



Describing with data

Unintended correlations.

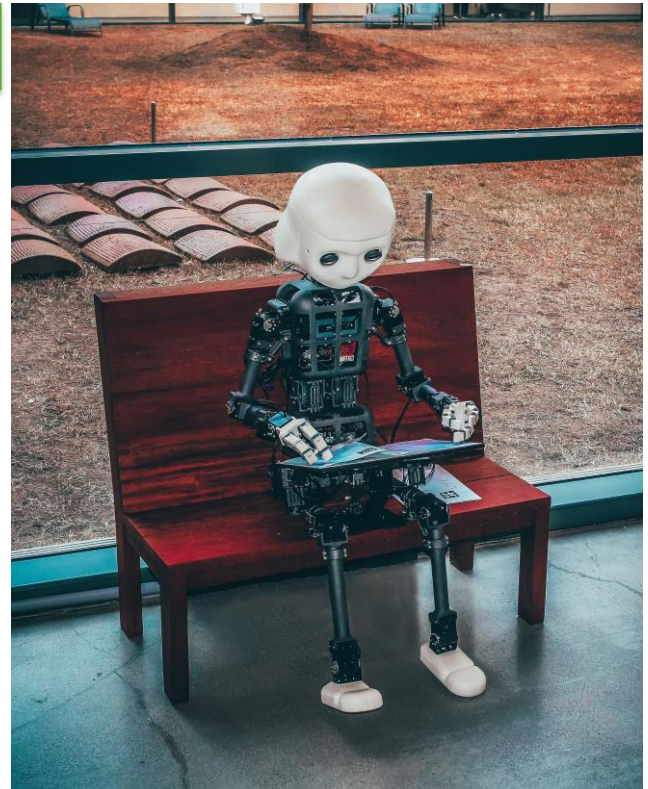
Recognizing a ruler instead of a cancer

January 11, 2020 Kerry Champion

[Skin cancer diagnosis AI](#) illustrates why it is important to provide explainability and to look for the connections between the model's behavior and actual causality operating in the system. By understanding why a prediction is being made and what a model says about the underlying system we can catch issues like this.

When dermatologists are looking at a lesion that they think might be

We give examples that we want the model to learn from, but we are also trying to be as efficient as possible. We don't want to describe everything in the data, we instead rely on the ability of machine learning to find the important features.

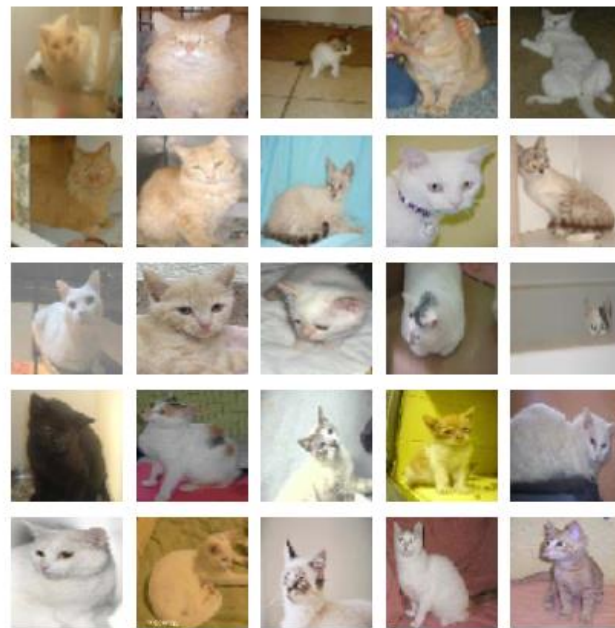


Describing with data

Unknown but important features.

Machine learning is taught to prefer the most likely answer. It may learn to be biased against unusual, but legitimate, cases.

We might not know that an AI is biased against a minority group, or an individual, until it's too late.



(a) Lowest scoring cats.

Colin Paterson, Radu Calinescu, and Chiara Picardi. "Detection and mitigation of rare subclasses in deep neural network classifiers." *2021 IEEE International Conference on Artificial Intelligence Testing (AITest)*. IEEE, 2021.

The people problem

By Elsa Maishman

BBC News

A robot broke a seven-year-old boy's finger during a chess match in Moscow last week, Russian news outlets report.

"The robot broke the child's finger," Sergey Lazarev, Moscow Chess Federation President, told Tass news agency. "This is of course bad."

Data from [Autonomous Cars, Robotaxis & Sensors 2022-2042](#) reveals for crashed vehicles that were operating in autonomous mode, 81 out of the 83 recorded incidents were caused by a human, either in another vehicle or as a misbehaving pedestrian.

Of 187 reports of autonomous vehicles accidents, just two could be attributed to the poor performance of the systems.

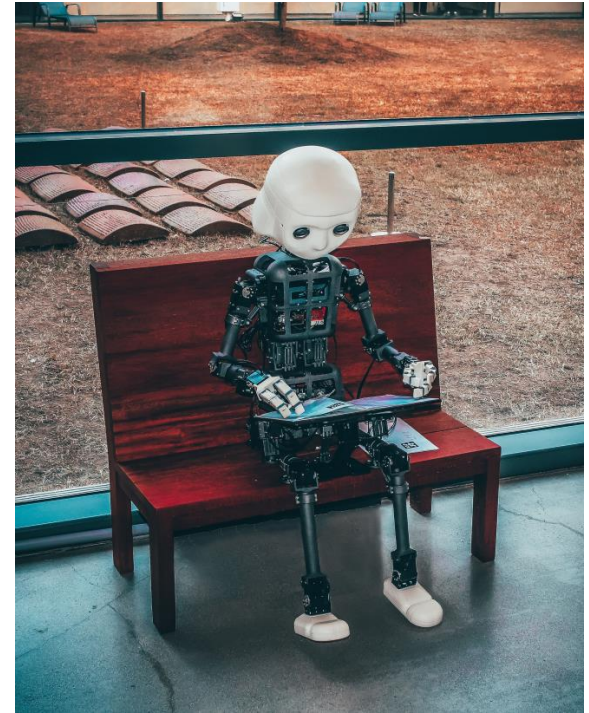


1. Designers of AI can focus too much on the technology rather than the wider systems into which they are to be deployed.
2. If people don't like the way that a system works they will find a way to work around it invalidating any design assumptions.

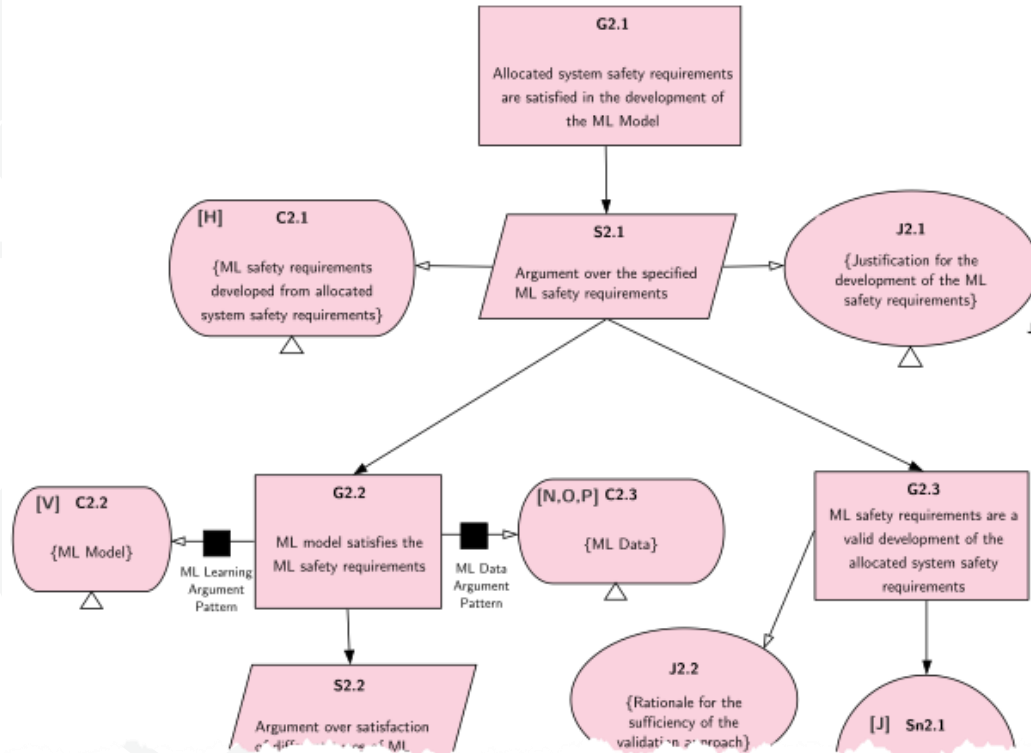
Assuring safety

Some points to note:

- Safety is not something that can be bolted on after the event.
- Safety is a systems level issue. We can not say a machine learnt algorithm is, in itself, safe.
- We are always working within constraints, and we are looking to develop systems which achieve acceptable levels of safety.



Assuring Safety



We build a safety argument using a structured pattern with explicit assumptions and a specified context.

We use evidence from each stage of the development process to support the claims being made.

Arguments may rely on sub claims and arguments.

When it all goes wrong

On March 18, 2018, a self-driving Uber car struck and killed a pedestrian crossing her bike in the middle of an Arizona roadway. At the steering wheel of the putative “autonomous vehicle,” a safety driver sat. Her job was to monitor the car’s systems and take over in the event of an emergency. The safety driver now may face criminal charges of vehicular manslaughter.

Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction, 2019, Madeleine Clare Elish.

There is a widely held belief that if we put a human in the loop then the system will be safe.

Advisory AI

Complexity	AI	Human	Outcome	Blame
Low	Correct	Agree	Success	None
High	Incorrect	Disagree	Success	None
High	Incorrect	Agree	Failure	?
High	Correct	Disagree	Failure	Human

↑
You wouldn't
know this?

↑
Causation?

*When would you
choose to disagree
with the AI?*

Explanation expected

Even if the AI is advisory, you may be expected to explain why you made your decision even if it's just accepting the AI's recommendation.

Was the decision wrong?

Just because a failure occurred does not mean the decision was wrong.

Can you explain the AI?

Is this a data problem, a training problem, a deployment problem ...

Is this a system problem?

Culture, training, reporting, reviewing, time pressures, ...

... the irony (is) that the more advanced a control system is, so the more crucial may be the contribution of the human operator.

Ironies of Automation, Lisanne Bainbridge, 1983

Sanctions

Analysis of responsibility attributions in complex, partly automated socio-technical systems has identified the risk that the nearest human operator may bear the brunt of responsibility for overall system malfunctions.

As we move towards integrating AI into healthcare systems, it is important to ensure that this does not translate into clinicians unfairly absorbing legal liability for errors and adverse outcomes over which they have limited control.

Clinicians Risk Becoming “Liability Sinks” for Artificial Intelligence: Lawton et al, 2023.



Who is to blame (%)?

AI developer

Regulator

Clinical team

Systems
integrator

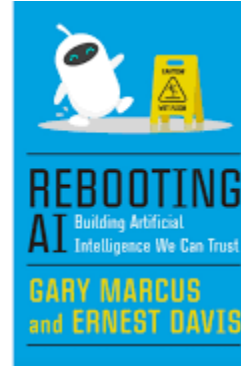
Employer

Clinician

Some useful resources



Hello World, How to be Human in the Age of the Machine: Hannah Fry.



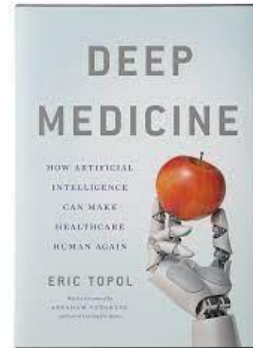
Rebooting AI: Building Artificial Intelligence We Can Trust: Gary Marcus and Ernest Davis

[dstl]

The Dstl Biscuit Book
Artificial Intelligence, Data Science and
(mostly) Machine Learning
1st edition revised v1.2



[Artificial Intelligence, Data Science and \(mainly\) Machine Learning: the Dstl Biscuit Book](#)



Deep Medicine. How Artificial Intelligence Can Make Healthcare Human Again: Eric Topol

Thank you

AI 'godfather' Yoshua Bengio feels 'lost' over life's work



One of the so-called "godfathers" of Artificial Intelligence (AI) has said he would have prioritised safety over usefulness had he realised the pace at which it would evolve.

[BBC News June 2023](#)