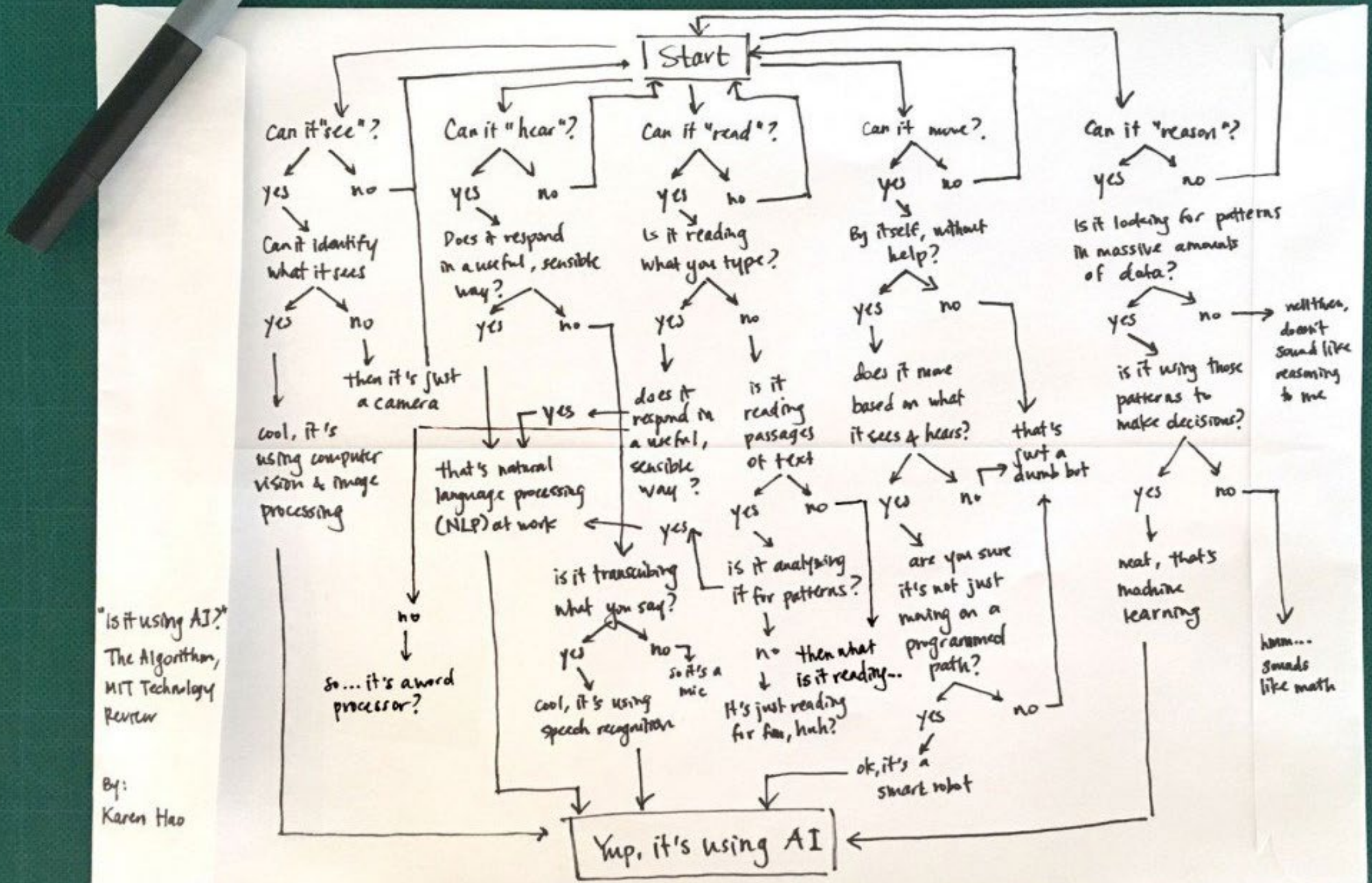
October 2019

Demonstrably doing AI accountability by design: problems, processes and tools

Peter Leonard
Principal, Data Synergies Pty Limited
Professor of Practice, UNSW Business School

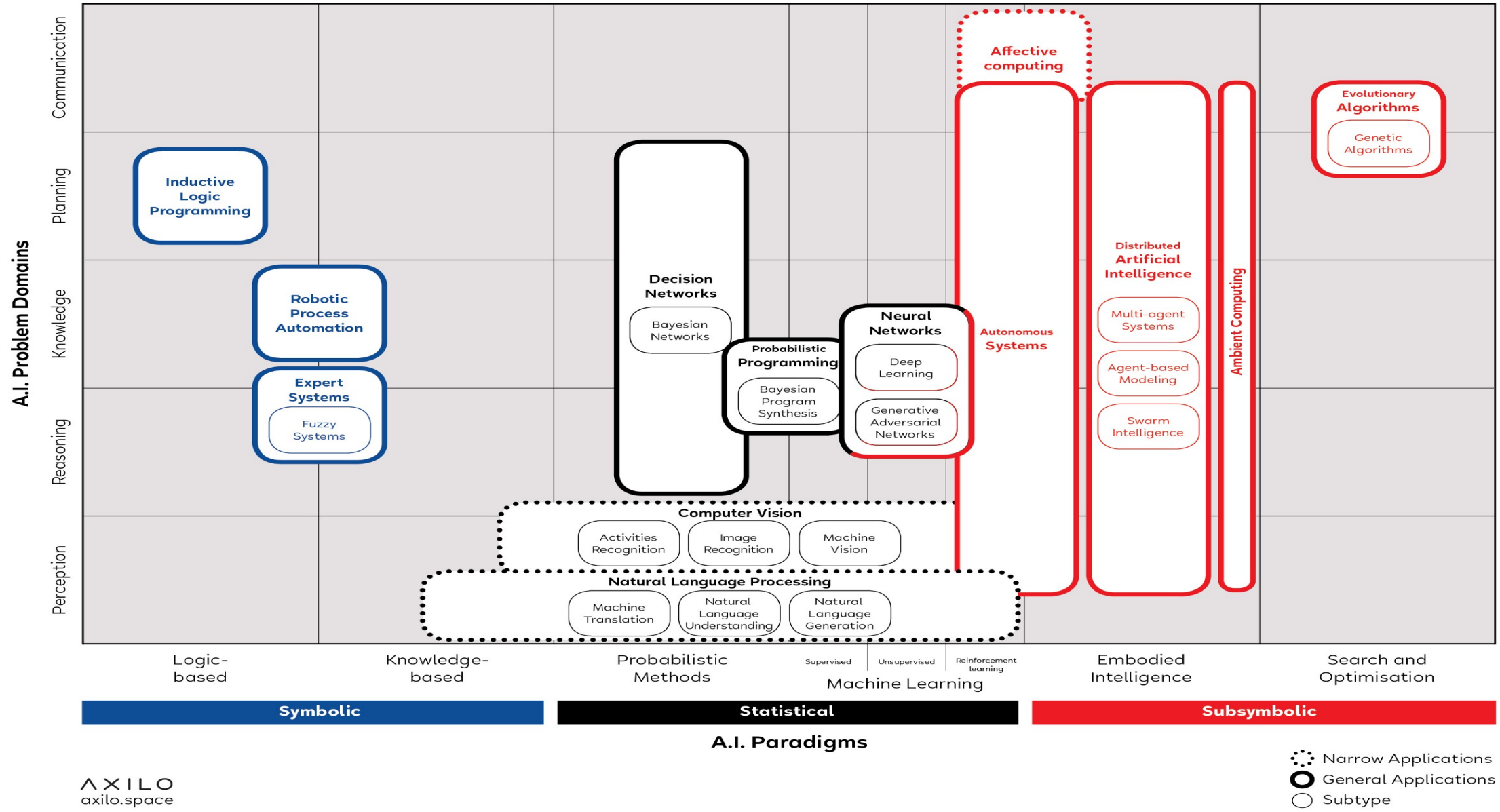
DATA SYNERGIES

DATA SYNERGIES



"Is it using AI?"
The Algorithm, MIT Technology Review

By:
Karen Hao



AXILO
axilo.space

DATA SYNERGIES

Algorithmic accountability BINGO

Trolley problem	Isaac Asimov's 3 laws of robotics	Apocalyptic doomsday scenario	Technological solutionism
Bias as something that exists outside of society	Namedropping of ethical theories (deontology, utilitarianism, virtue ethics)	"We need more oversight!" (But we don't know how..)	Algorithmic accountability not drawing on accountability theory
Self-driving cars	GDPR (as savior)	Call for ethical classes for data scientists/engineers	China's AI efforts
Hippocratic oath for data scientists	Robots everywhere!!?!	Right to explanation (and the belief that that will fix everything)	Predictive policing/parole recommendation

By Maranke Wieringa

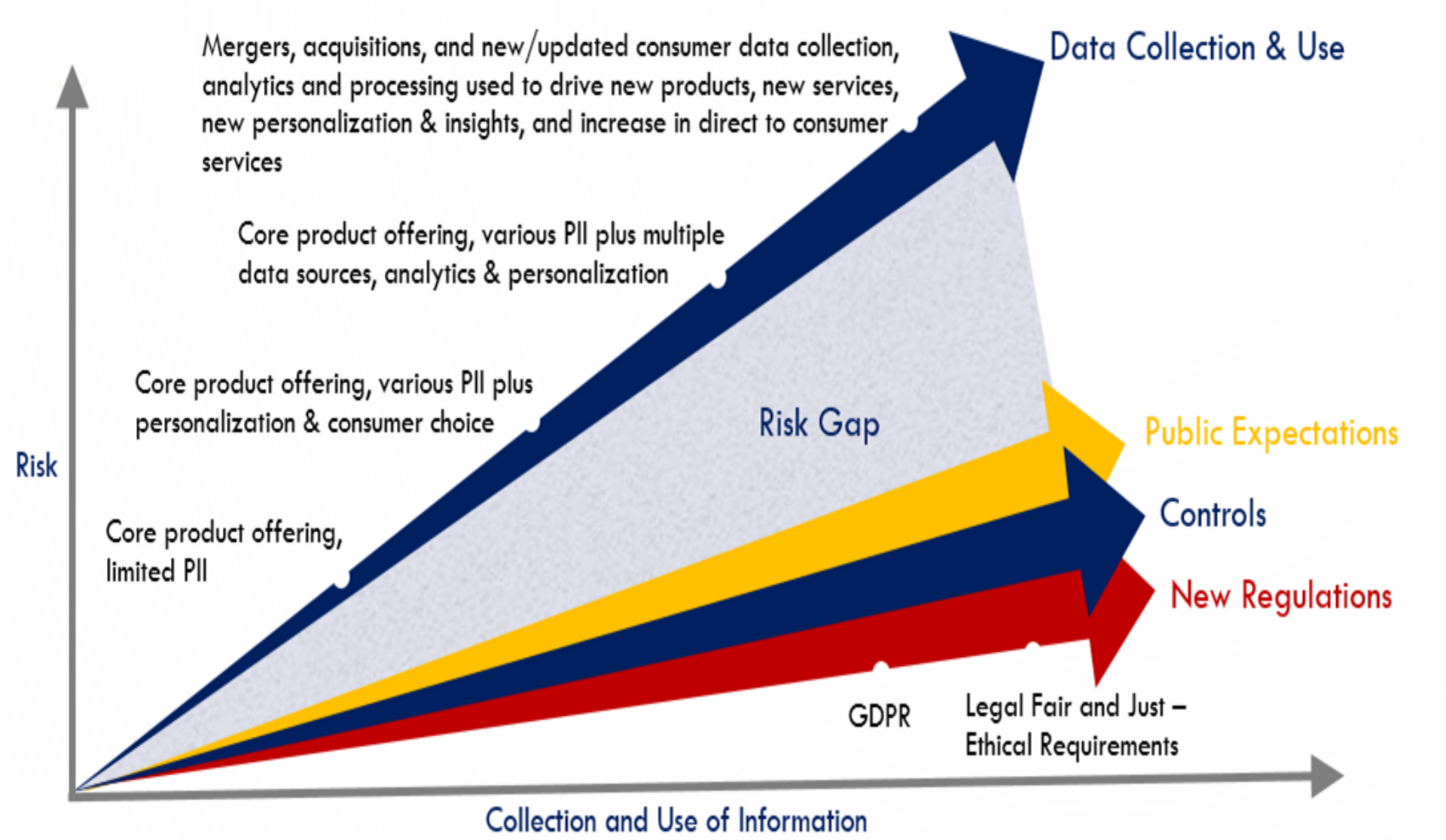
MACHINE LEARNING

Do Stupid Things Faster with More Energy

Smartphones and IOT

Data Issues Cubed

Data as business driver and risk



Source: Peter Cullen and Information Accountability Foundation February 2018

DATA SYNERGIES

Deficit in digital trust

- Someone is making money from ‘my data’, and I’m not
- I’m powerless and being gamed, so I click-through
- Cambridge Analytica, *taking [insert] seriously*, and FB billboarding the world
- ABS and ‘Census fail’
- surveillance and tracking
- MyHealth Record opt-in to opt-out
- protections re secondary uses of MyHealth Record data
- non-digital - institutional abuses: banks, super funds, vulnerable kids, vulnerable elderly…

DATA SYNERGIES

Are we barking mad(ly)?

- AI is not some ‘thing’ that is fundamentally different.
- AI is an application of advanced data analytics, sometimes presenting further tricky issues by taking humans out-of-the-loop, or by being a black box with no explainability.
- Mostly, AI presents similar practical issues as to transparency and accountability by design issues as the issues that we should already be grappling with in relation to here and now digital applications and technologies, including:

  smartphones,

  personal IoT - personal wellness devices and other health IoT applications,

  IoT applications where the affected individual doesn’t know they are observed or sensed;

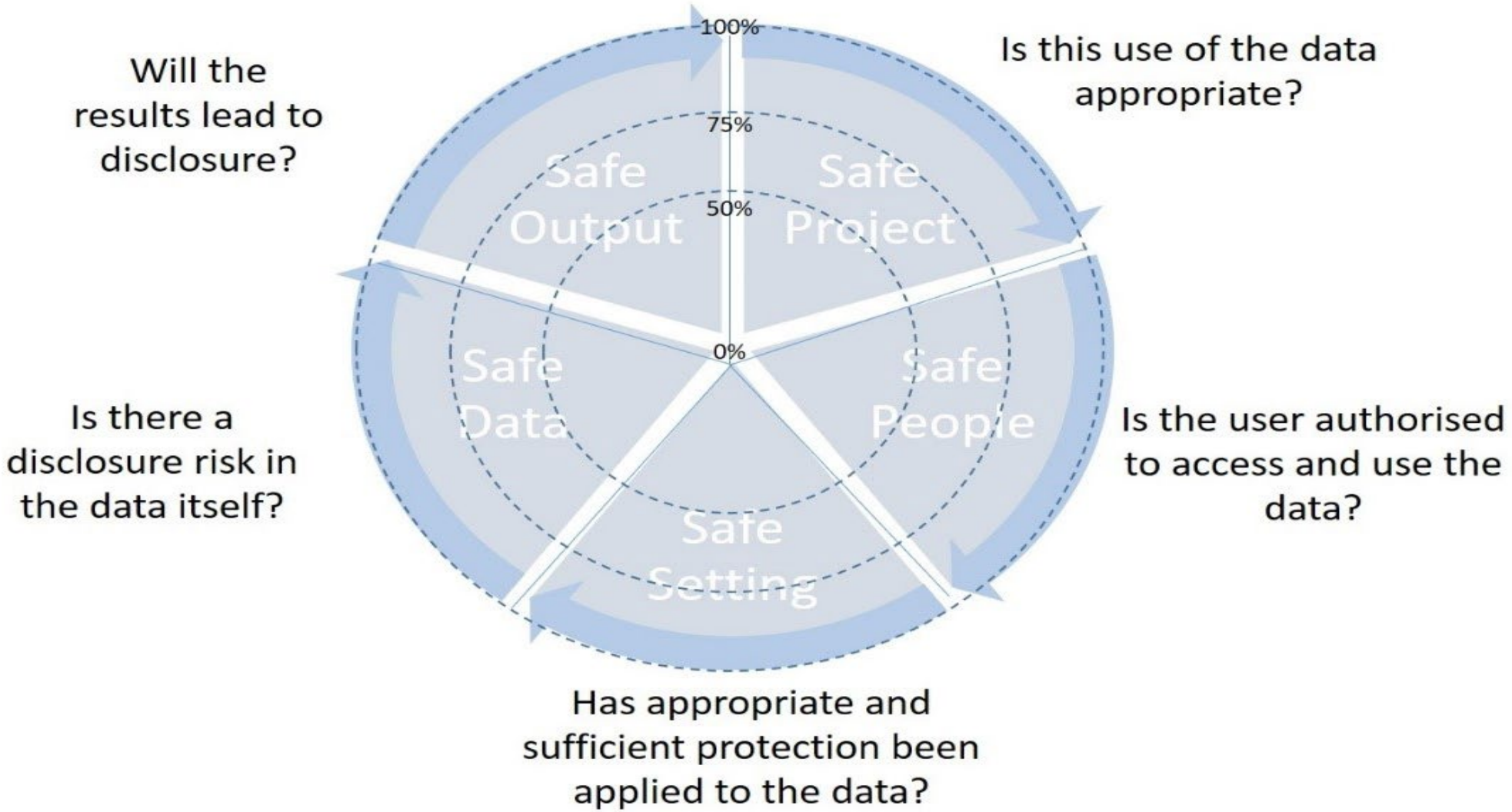
  the MyHealth Record system

  many health data linkage and data sharing initiatives.
- Susceptibility to common viruses, AI aversion and AI deference.
- Deference to ÁI and data science is really risky, but so is (any) human in the loop
- Need to be really clear about risks: specifically, uses and applications of data relating to humans in ways that affect those humans, or other humans, or our environment.

DATA SYNERGIES

Filling the gap through governance

- Explainability: the Brian Cox Test
- Sustainability and fly-by-night havoc
- Statements and principles – ethics washing and rights washing
- Agency: click-throughs, responsibility and blame shifting
- Causes: incentives and sanctions
- Enablers: caring and impact assessment capabilities (governance + methodologies + tools + processes)
- FEAT/FATE: fair, equitable and explainable, accountability and transparent
- Outputs and outcomes: who's responsible for what, and why them?



Source: Australian Computer Society data sharing White Paper (forthcoming, Oct 2019

RESPONSIBLE AI MADE EASY

FOR ORGANISATIONS

Using Artificial Intelligence (AI) in your company?

Help your customers understand and build their confidence in your AI solutions.

PRINCIPLES FOR RESPONSIBLE AI

- **DECISIONS MADE BY AI SHOULD BE** EXPLAINABLE, TRANSPARENT AND FAIR
- **AI SYSTEMS, ROBOTS AND DECISIONS SHOULD BE** HUMAN-CENTRIC

4 FACTORS TO CONSIDER

INTERNAL GOVERNANCE STRUCTURES & MEASURES

- Clear roles and responsibilities in your organisation
- SOPs to monitor and manage risks
- Staff training

RISK MANAGEMENT IN AI DECISION-MAKING

- Appropriate degree of human involvement
- Minimise the risk of harm to individuals

OPERATIONS MANAGEMENT

- Minimise bias in data and model
- Regular reviews of data and model

CUSTOMER RELATIONSHIP MANAGEMENT

- Make AI policies known to users
- Allow users to provide feedback, if possible
- Use simple language

FIND OUT MORE ABOUT THE PDPC'S MODEL AI GOVERNANCE FRAMEWORK AT **WWW.PDPC.GOV.SG/MODEL-AI-GOV.**

An initiative by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY pdpc PERSONAL DATA PROTECTION COMMISSION SINGAPORE

In support of: SG:D EMPOWERING POSSIBILITIES

DEGREE OF HUMAN INVOLVEMENT

Determine the degree of human involvement in your AI solution that will minimise the risk of adverse impact on individuals.

SEVERITY OF HARM

LOW — HIGH

Human-out-of-the-loop
AI makes the final decision without human involvement, e.g. recommendation engines.

Human-over-the-loop
AI decides but the user can override the choice, e.g. GPS map navigations.

Human-in-the-loop
User makes the final decision with recommendations or input from AI, e.g. medical diagnosis solutions.

HUMAN INVOLVEMENT: HOW MUCH IS JUST RIGHT?

EXAMPLE

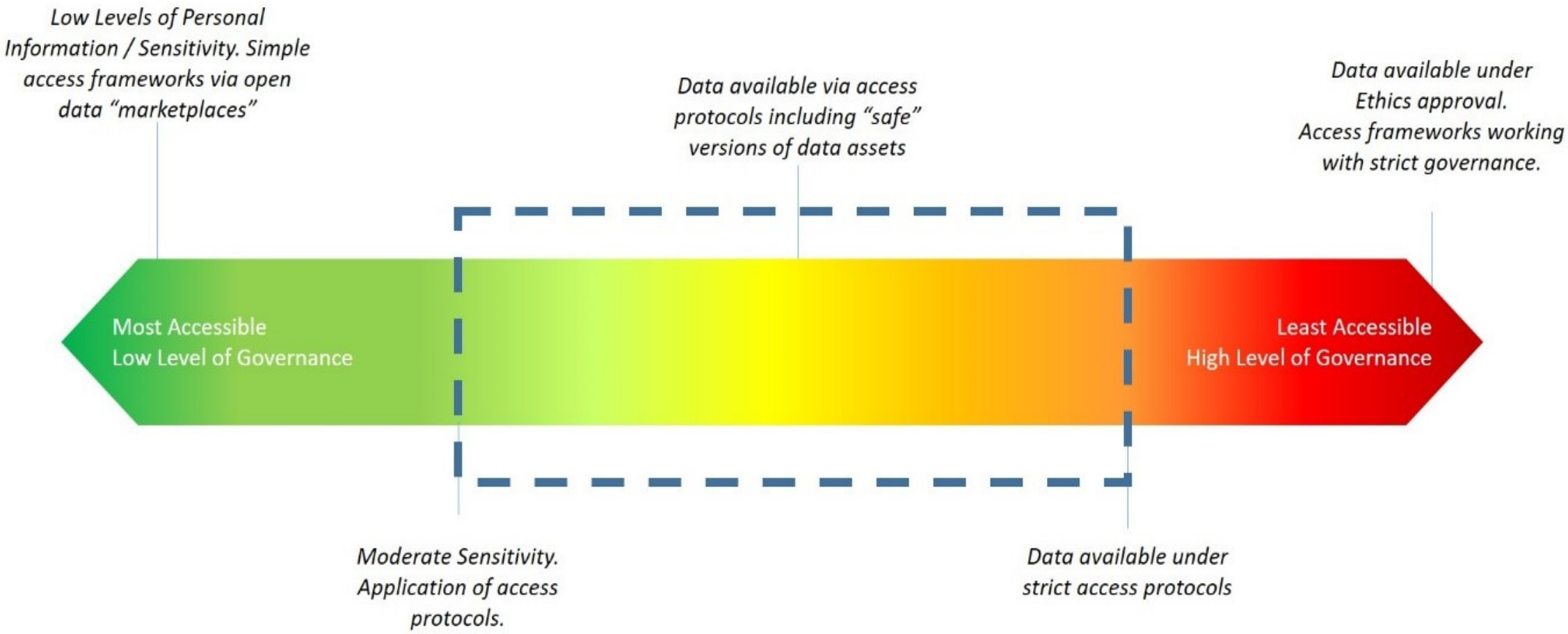
An online retail store wishes to use AI to fully automate the recommendation of food products to individuals based on their browsing behaviours and purchase history.

What should be assessed?

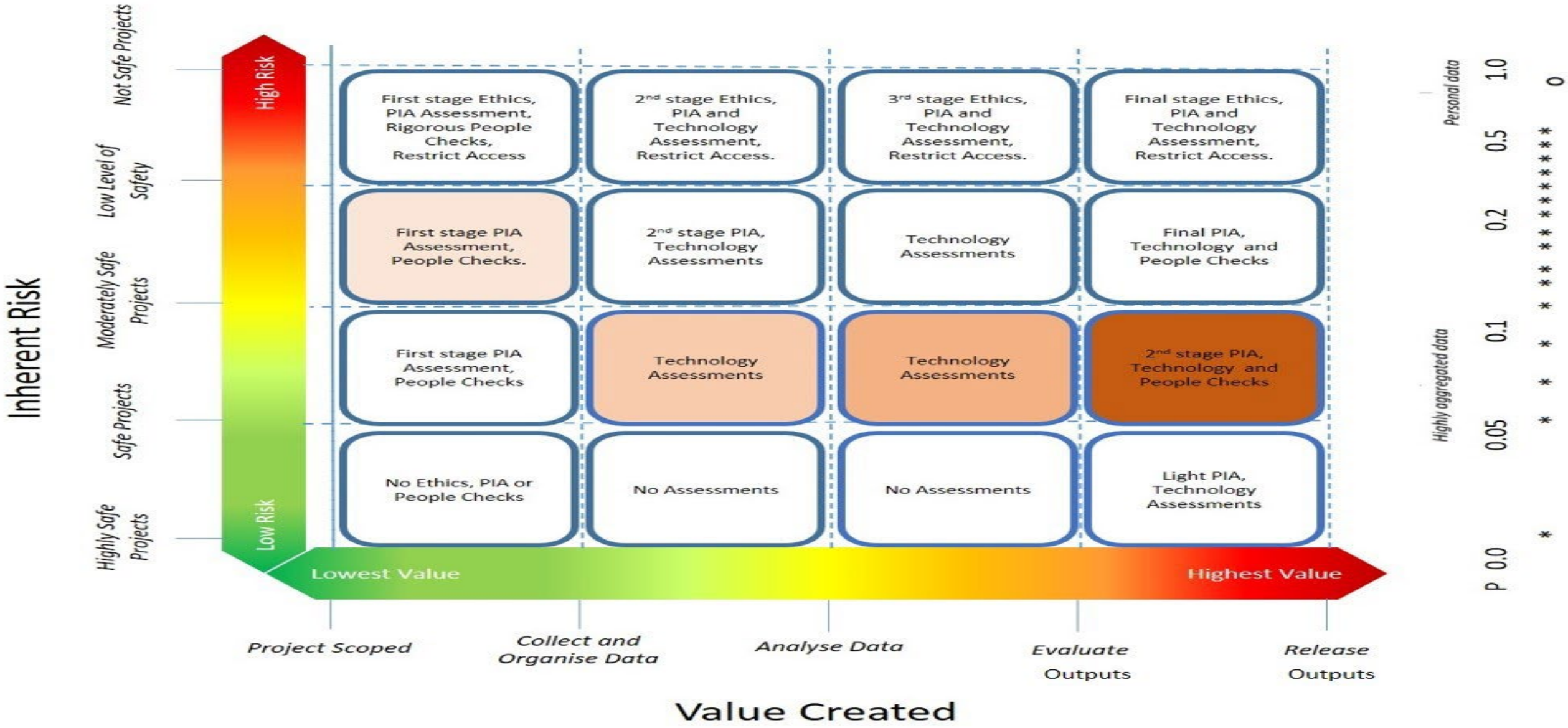
What is the harm?
One possible harm could be recommending products that the customer does not need or want.

Is it a serious problem?
Wrong product recommendations would not be a serious problem since the customer can still decide whether or not to accept the recommendations.

Recommendation:
Given the low severity of harm, the human-out-of-the loop model could be considered for adoption.



Source: Australian Computer Society data sharing White Paper (forthcoming, Oct 2019



DATA SYNERGIES

Source: Australian Computer Society data sharing White Paper (forthcoming, Oct 2019

Data Ethics Canvas

Data sources

Name/describe key your project's data sources, whether you're collecting data yourself or accessing via third parties.

Is any personal data involved, or data that is overwise sensitive?

Limitations in data sources

Are there limitations that could influence your project's outcomes?

Consider:

> bias in data collection, inclusion/exclusion, analysis, algorithms

> gaps or omissions in data

> provenance and data quality

> other issues affecting decisions, such as team composition

Sharing data with others

Are you going to be sharing data with other organisations? If so, who?

Are you planning to publish any of the data? Under what conditions?

Ethical and legislative context

What existing ethical codes apply to your sector or project? What legislation, policies, or other regulation shape how you use data? What requirements do they introduce?

Consider: the rule of law; human rights; data protection; IP and database rights; anti-discrimination laws; and data sharing, policies, regulation and ethics codes/frameworks specific to sectors (eg health, employment, taxation).

Rights around data sources

Where did you get the data from? Is it produced by an organisation or collected directly from individuals?

Was the data collected for this project or for another purpose? Do you have permission to use this data, or another basis on which you're allowed? What ongoing rights will the data source have?

Your reason for using data

What is your primary purpose for collecting and using data in this project?

What are your main use cases and business model?

Are you making things better for society? How and for whom?

Are you replacing another product or service as a result of this project?

Communicating your purpose

Do people understand your purpose – especially people who the data is about or who are impacted by its use?

How have you been communicating your purpose? Has this communication been clear?

How are you ensuring more vulnerable individuals or groups understand?

Positive effects on people

Which individuals, groups, demographics or organisations will be positively affected by this project? How?

How are you measuring and communicating positive impact? How could you increase it?

Negative effects on people

Who could be negatively affected by this project?

Could the way that data is collected, used or shared cause harm or expose individuals to risk of being re-identified? Could it be used to target, profile or prejudice people, or unfairly restrict access (eg exclusive arrangements)?

How are limitations and risks communicated to people? Consider: people who the data is about, people impacted by its use and organisations using the data.

Minimising negative impact

What steps can you take to minimise harm?

What measures could reduce limitations in your data sources? How are you keeping personal and other sensitive information secure?

How are you measuring, reporting and acting on potential negative impacts of your project?

What benefits will these actions bring to your project?

Engaging with people

How can people engage with you about the project?

How can people affected correct information, appeal or request changes to the product/service? To what extent?

Are appeal mechanisms reasonable and well understood?

Openness and transparency

How open can you be about this project? Could you publish your methodology, metadata, datasets, code or impact measurements?

Can you ask peers for feedback on the project? How will you communicate it internally?

Will you publish your actions and answers to this canvas openly?

Ongoing implementation

Are you routinely building in thoughts, ideas and considerations of people affected in your project? How?

What information or training might be needed to help people understand data issues?

Are systems, processes and resources available for responding to data issues that arise in the long-term?

Reviews and iterations

How will ongoing data ethics issues be measured, monitored, discussed and actioned?

How often will your responses to this canvas be reviewed or updated? When?

Your actions

What actions will you take before moving forward with this project? Which should take priority?

Who will be responsible for these actions, and who must be involved?

Will you openly publish your actions and answers to this canvas?

Wheat and chaff, good and bad

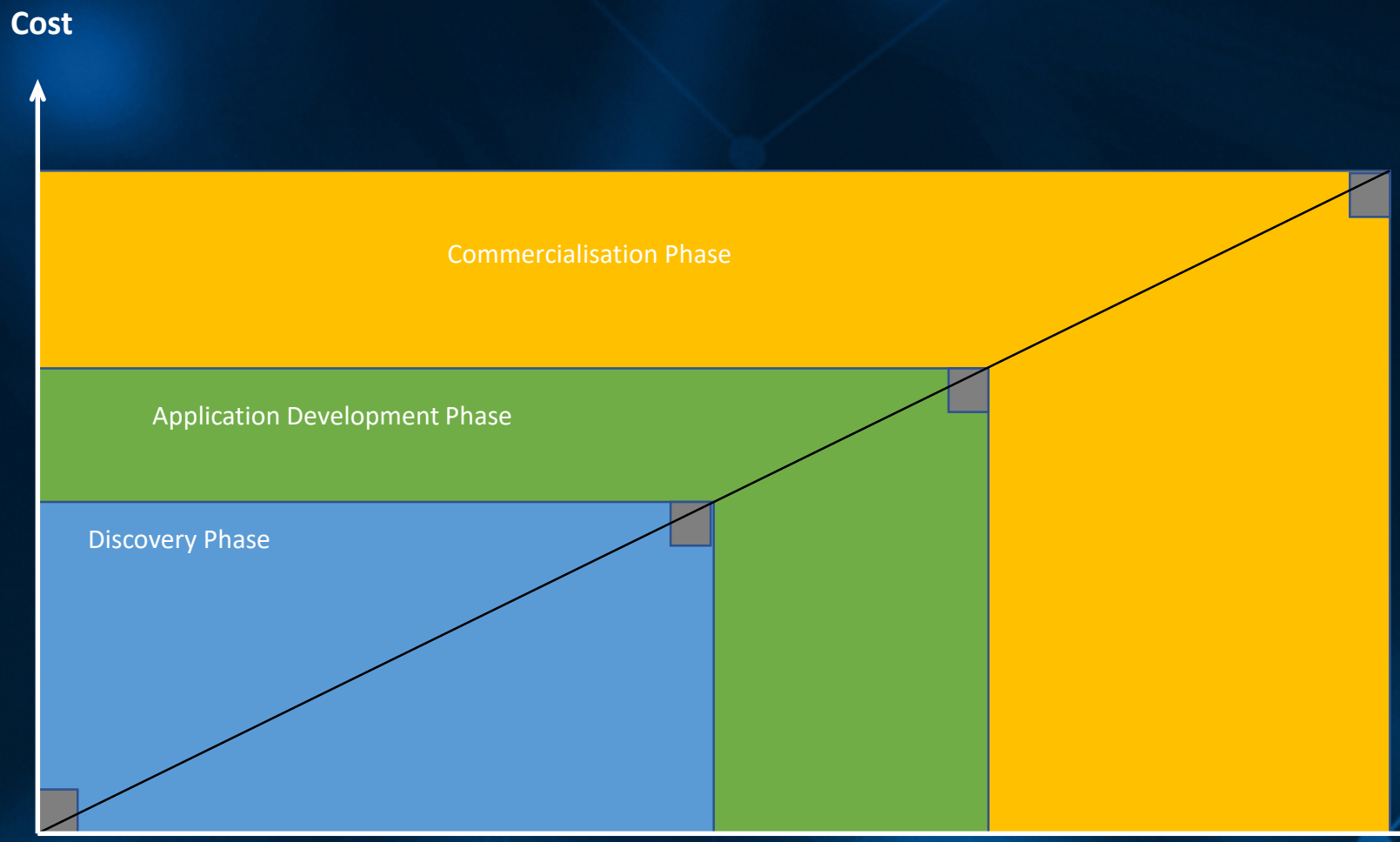
- owners, custodians and stewards:
  asset and liability
- observed data and 'digital exhaust‘
- volunteered, transformed, curated and inferred data

- ‘data isn't oil‘ - harvesting v value creation
- the human factor
  - transformation and insights
  - algorithms, processes and methodologies

Is data your best asset (that you never own)?

- rights and duties of custodians and stewards
- end-to-end data governance
- contracts, trade secrets, IP ownership, privacy and trust
- relative confidentiality
- differential privacy, labs and clean rooms
- supply and demand side data ecosystems
- (unilateral) contracts

Data Analytics Project Review Framework



Source: Data Synergies January 2019

Gate 1
Data Impact Assessment
Ethics Review (if Human Research)

Gate 2
Outputs (Agile) Assessment
Revise DIA + ER

Gate 3
Outcomes Assessment, check DIA + ER

Post-Imp Review
Feedback and improvement