BEHAVIOURAL SCIENCE MEETS AI

A PRIMER

ANNA, ANNABEL, ISABEL, MUSKAAN

WHY REVIEW THE INTERSECTION OF AI & BEH SCI?

There is much discussion about the intersection between AI and Behavioural Science but a lack of detailed research in this interdisciplinary area and a lack of clarity and consensus about the current state of play, the key issues, and the direction of future research travel.

This primer is intended to:

- ✓ ***Synthesise*** the current state of play and key definitions and categorizations of AI
- ✓ ***Identify initial applications*** *and intersection between AI and BehSci*
- ✓ *Highlight areas* of **challenges** and **future research**.

AGENDA

This presentation will offer 3 sections:

Part one – the big picture

- Definitions & Terms
- Framework of Relevant types of AI
- Applications of AI & Behavioural Science

Part two – the key challenges

- Ethical Considerations
- Regulation

Part three – recommendations and next steps

PART ONE:
THE BIG PICTURE

CATEGORIZATION & TYPES OF ARTIFICIAL INTELLIGENCE

"A computer, or a system, made up of algorithms that is capable of intelligent behaviour" (Burns et al., 2022)

"An umbrella term for a range of algorithm-based technologies that solve complex tasks by carrying out functions that previously required human thinking" (Bergin & Tannock, 2020)

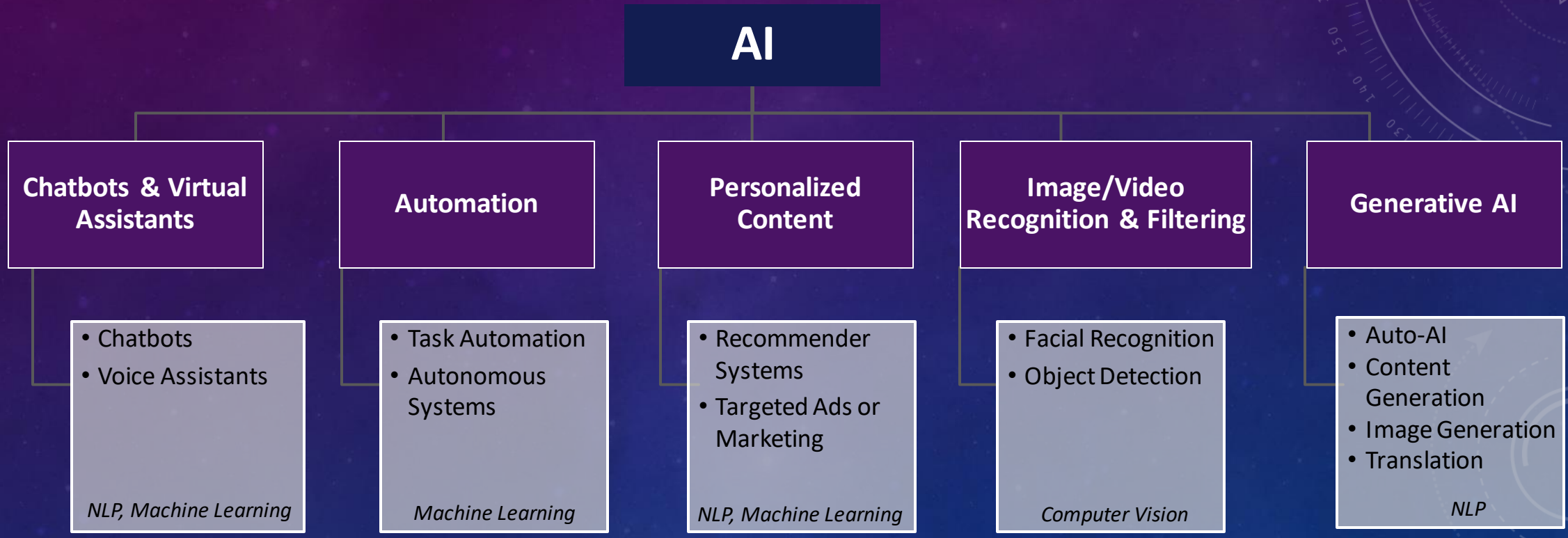
"The capability of a machine to imitate intelligent human behavior" (Marr, 2018)

What is "Intelligence"

What is "General" versus "Narrow" AI?

Is AI a field of study or a capability?

RELEVANT TYPES OF AI



HUMAN IN THE LOOP*

High / All AI

Level of Human in the Loop

High / All Human

Fully Automated	Semi-Automated	Assisted	Advisory	Dependent
• Autonomous pilot systems • Robotic Process Automation • Self-driving cars	• Recommender Systems • Predictive Analytics • Chatbots	• Natural Language Processing (NLP) • Computer Vision • Speech Recognition	• Data Analysis • Business Intelligence	• Labeling tasks for Supervised Learning • Quality Control Systems • Interactive Voice Response (IVR) Systems

*Initial Mapping, Work In Progress (WIP)

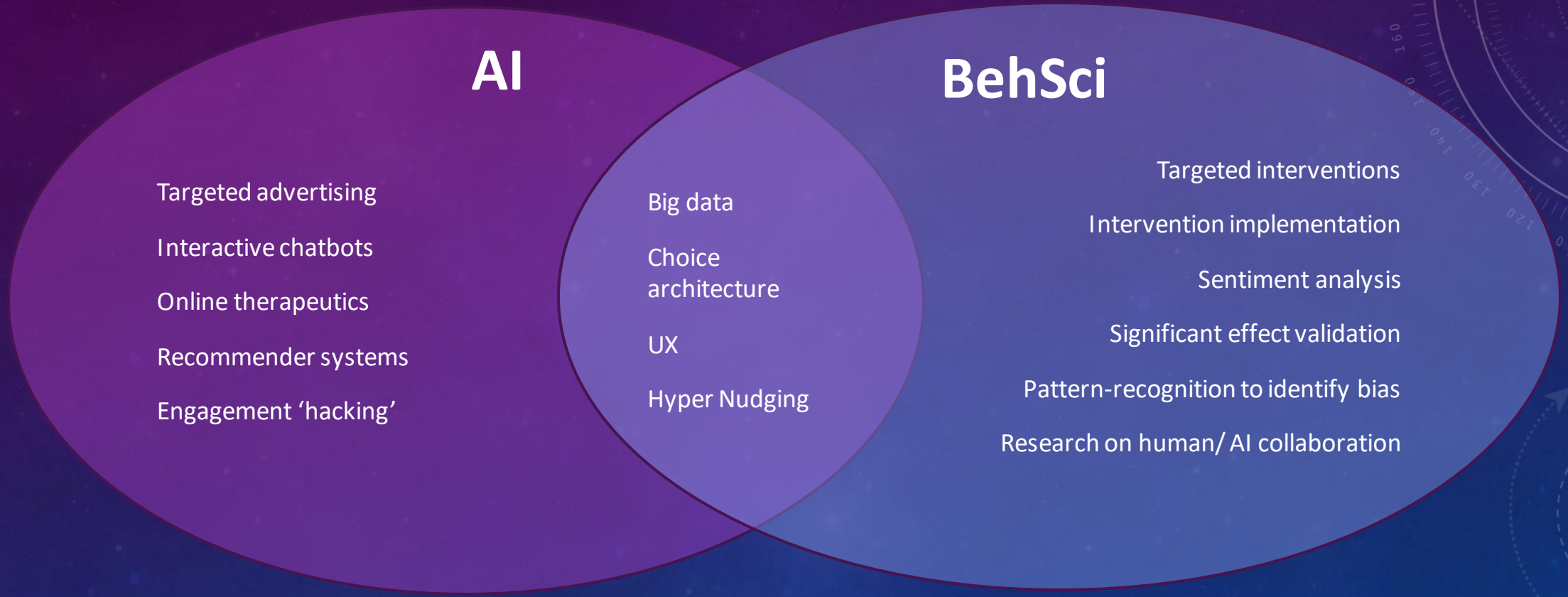
INTERSECTION OF AI & BEHAVIOURAL SCIENCE

APPLICATIONS OF AI & BEHAVIOURAL SCIENCE

"algorithmic nudging is much more powerful than its non-algorithmic counterpart. With so much data about workers' behavioral patterns at their fingertips, companies can now develop personalized strategies for changing individuals' decisions and behaviors at large scale. These algorithms can be adjusted in real-time, making the approach even more effective."

Algorithmic nudges don't have to be unethical" Marieke Mohlmann, HBR, April 2021

HOW AI & BEHAVIOURAL SCIENCE INTERRELATE



Applications

- **AI as a tool in BehSci**
- **Machine Behaviour**
- **Big Data <> Behaviour**
- **BehSci as a Tool in AI Development**

AI AS A TOOL IN BEHSCI

The extent to which AI extends the quality of behavioural research, insights, and interventions; Leveraging various forms of Artificial Intelligence to enhance behavioural science in practice, research, gathering insights, etc.

1. **Improving data collection (e.g. emotional recognition systems)**

2. **Recognising heterogenous causal effects from false positives**

HB CP Human Behaviour Change Project

1. Knowledge systems
2. Synthesis
3. Information extraction

Modelling behavioural responses in an 'artificial society'

BEHSCI AS A TOOL IN AI DEVELOPMENT AND INTERFACES

The usage of behavioural science theory, principles, and findings in the development and design of AI tools, systems, and agents;
The extent to which behavioural insights can bridge the gap between emotional intelligence and AI

EXAMPLE USE CASES

Identification of biases to guide the creation of humane AI

- Accuracy (e.g. debiasing hiring algorithms)
- Interactivity (e.g. social humanoid robots)

Developing Explainability of (XAI) AI

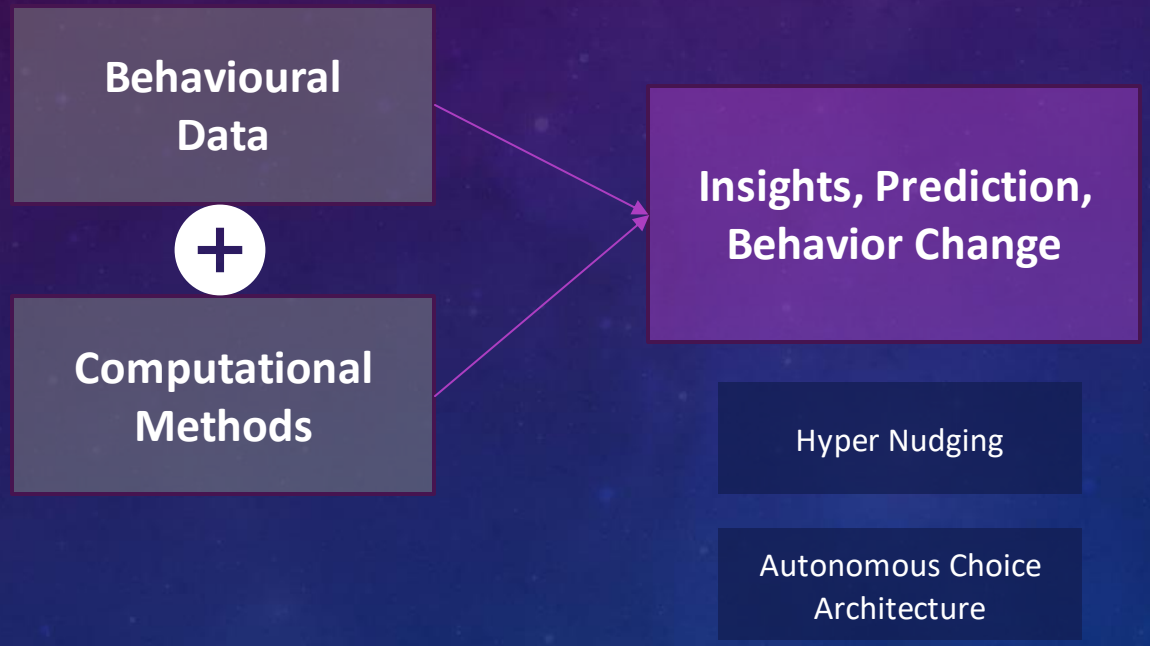
- How exactly does input become output
- May aid in regulation how how AI provides conclusion / advice

Design of AI Interface

- Transparency in decision-making processes to improve user trust
- Anthropomorphization of Agent / Interface

BEHAVIORAL DATA SCIENCE

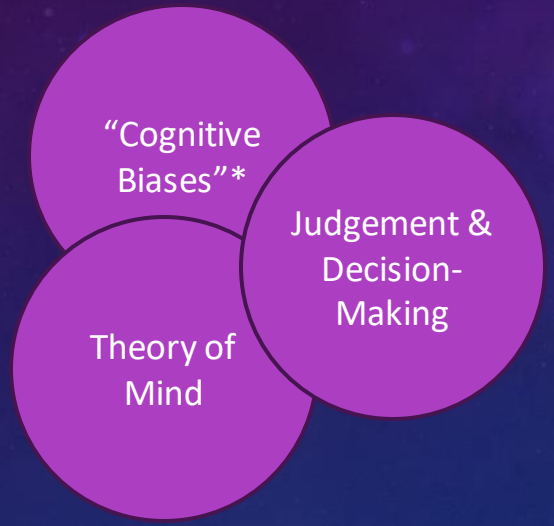
A combination of behavioural science techniques and computational approaches to predict, uncover insights on, (and change) human behavior. It rests on the combination of big data with a deep understanding of human behaviour/behavioural insights



"Behavioral data science tries to learn more about human psychology, its ideas, and its biases so that better predictions can be made, or patterns can be found." (Behera, 2023)

MACHINE BEHAVIOUR

Research to understand, assess, and uncover the patterns of 'decision-making in Artificial Intelligence, especially Generative AI and chatbots to assess the extent these 'machines' exhibit patterns of thinking similar to humans.



Select Research Highlights

Using cognitive psychology to understand GPT-3	Assessed GPT-3's performance on a variety of vignette-based tasks from the cognitive psychology literature
Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods	Evaluation on the usage of different sub-fields of psychology to assess machine behaviour, focusing on different lenses and use cases for each
Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?	Evaluated GPT-3's qualitative presentation of various 'cognitive biases' using well studied vignettes from behavioural science

Both private sector and academia have a stake in this application and leading companies such as Microsoft have funded Machine Behavior studies

PART TWO:
KEY CHALLENGES

ETHICAL FRAMEWORKS HAVE BEGUN TO EMERGE

OECD AI principles (May 2019)

Inclusive, sustainable, wellbeing
Fair & human-centred
Transparent & explainable
Robust, Safe & Secure
Accountability

Behavioural Science "FORGOOD" (Lades and Delaney, 2022)

***F**airness*
***O**penness*
***R**espect*
Goals
Opinions
Options
***D**elegation*

Is it clear when to apply which framework? Might 'neither' be the result
No requirement to follow either or accountability for shortfall - transparency

MIND THE GAP

*BUT issues arise at the intersection that **are not explicitly addressed by either framework***

Data privacy inc biometric
Bodily autonomy
Addiction
Real time interventions
Hypernudging
Algorithmic feedback loops
Manipulation
Dark nudges/ sludges
Misrepresentation by anthropormophism
New unprotected platforms

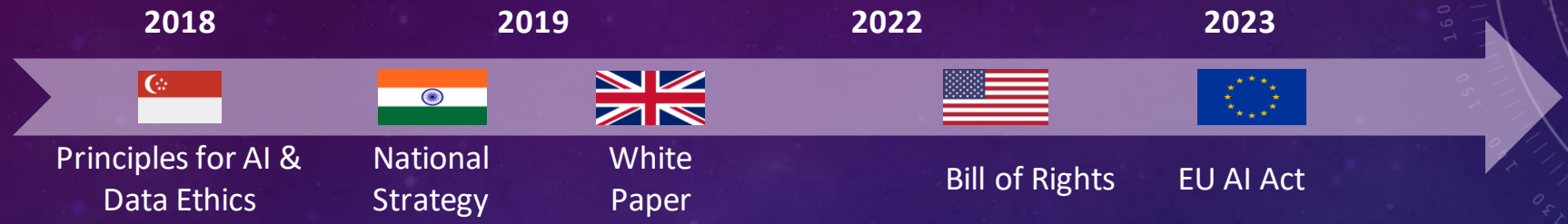
These applications may not have been envisaged in the design of either framework

The combination creates new ethical challenges:

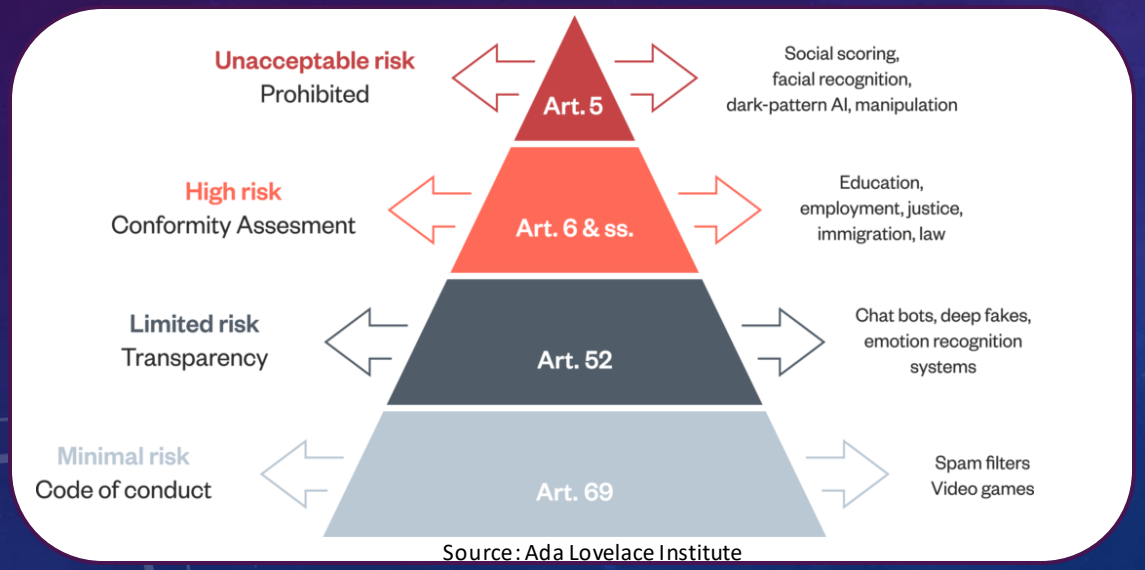
- Certain issues only arise at the intersection
- Existing issues that become turbo charged or qualitatively different (e.g., privacy)

REGULATION

AI Regulation Timeline



Rules-based approach (EU)



Source : Ada Lovelace Institute

Principles-based approach

Safety and robustness	US, UK
Protection against discrimination/ Fairness	US, UK, Singapore
Privacy	US
Transparency and explainability	US, UK, Singapore
Human alternatives, testing, and redress	US, UK
Accountability and governance	UK, Singapore

PART THREE:
RECOMMENDATIONS & NEXT STEPS

RECOMMENDATIONS AND NEXT STEPS*

Opportunities

AI can turbo charge BehSci

- Opportunities for self-optimization
- More effective nudging, more targeted design, more reliable findings

BehSci is turbo charging AI (and is already embedded in digital marketing & design)

- Personalized nudging, Social robots, embodied AI, moral agency, emotional recognition, environmental cost,

Outstanding Considerations

The power of the combination also materializes in risks and ethics

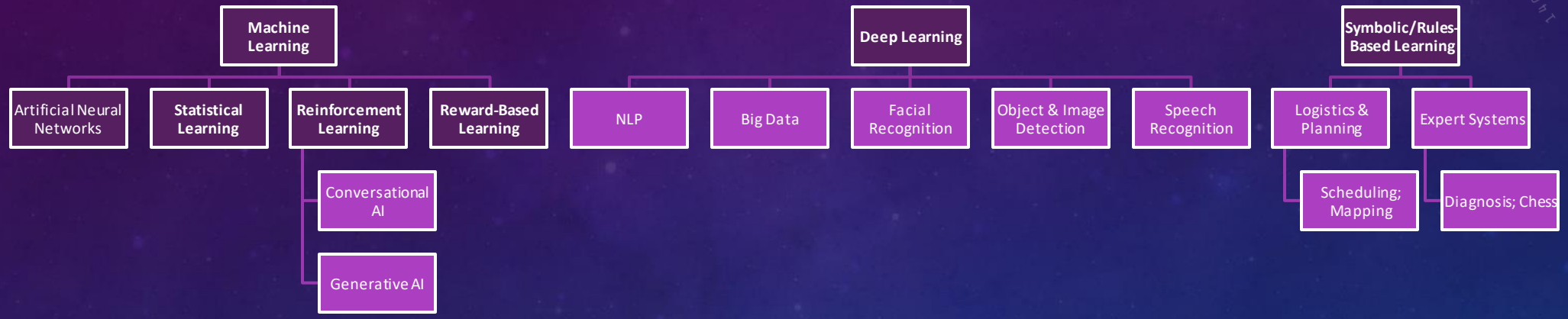
- Autonomous morality, emotional recognition, environmental cost, embodied AI

Where does Behavioural Science belong in AI research and development?

- Is this a new discipline, are they converging, how to navigate the gaps?

**Initial thoughts, non exhaustive*

APPENDIX


Machine Learning
Artificial Neural Networks
Statistical Learning
Reinforcement Learning
Reward-Based Learning
Conversational AI
Generative AI
Deep Learning
NLP
Big Data
Facial Recognition
Object & Image Detection
Speech Recognition
Symbolic/Rules-Based Learning
Logistics & Planning
Expert Systems
Scheduling; Mapping
Diagnosis; Chess

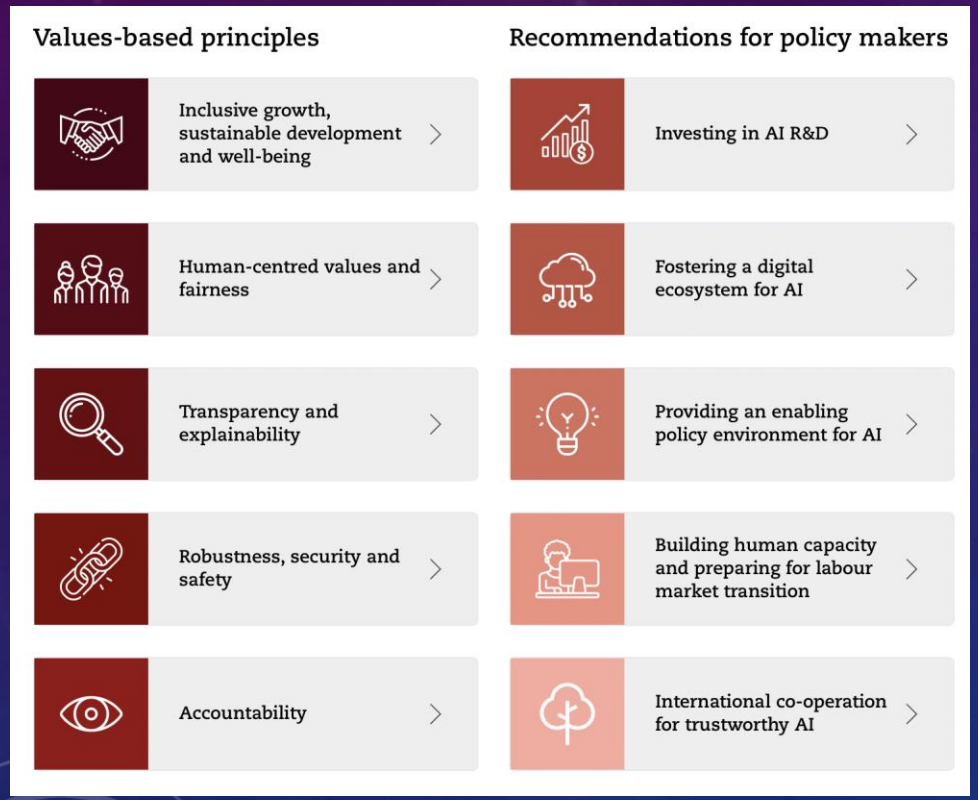
INTEGRATION OF AI

Level of Integration

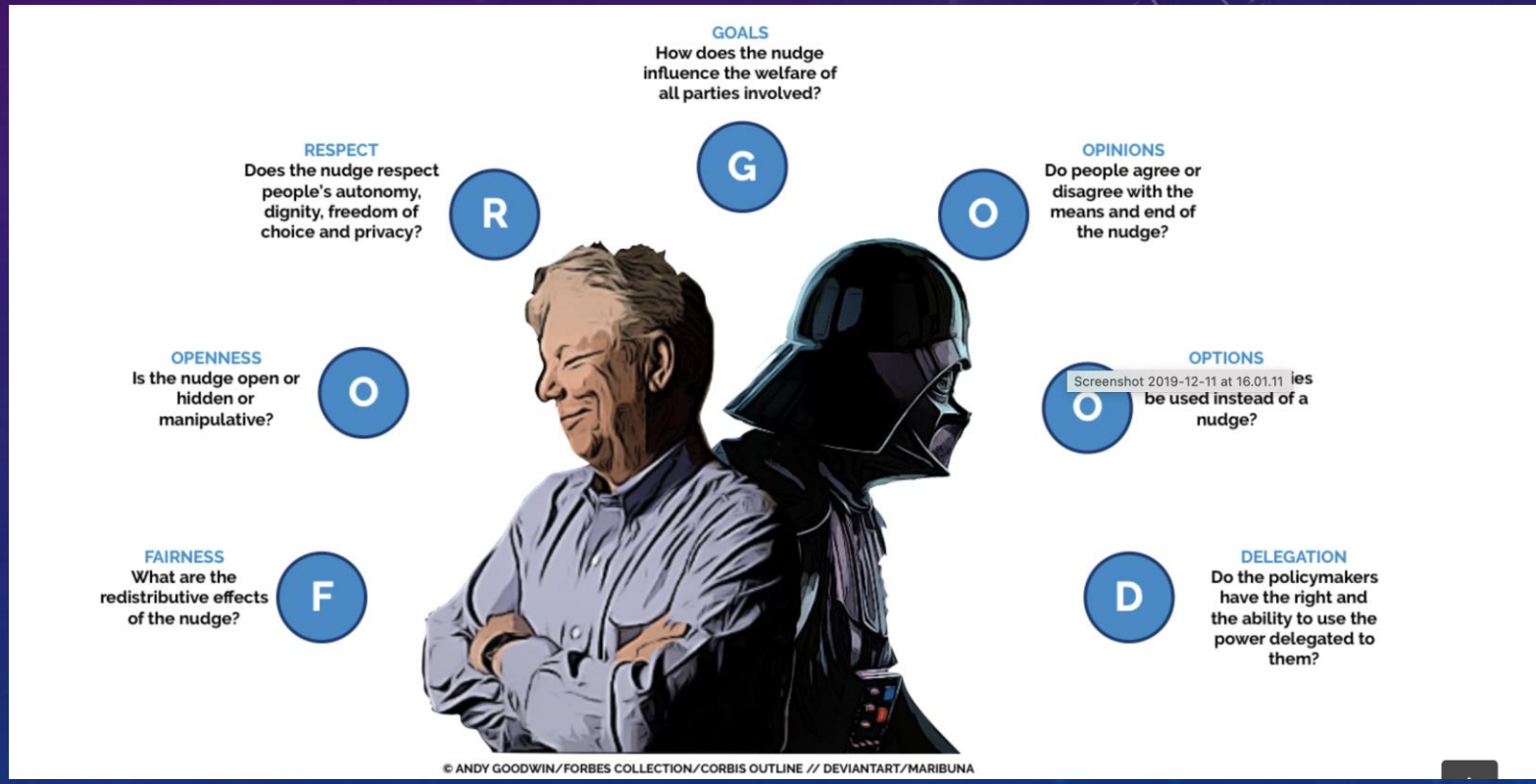
Integration	Description	Domains	Sub-types
Fully Automated	These are AI systems capable of operating without human intervention. They make decisions based on pre-set algorithms and can learn from their own experiences.	Autonomous pilot systems, Robotic Process Automation, Self-driving cars	Supervised Learning, Reinforcement Learning, Deep Learning
Semi-Automated	AI systems that require some degree of human input or intervention. They support human decision-making but don't replace it.	Recommender Systems, Predictive Analytics, Chatbots	Semi-Supervised Learning, Interactive Learning
Assisted	These AI systems provide help and augmentation to human operators, but the final decision-making still lies with humans. They provide analysis, recommendations, or insights to aid decision-making.	Natural Language Processing (NLP), Computer Vision, Speech Recognition	Transfer Learning, Multimodal Learning
Collaborative	Collaborative AI systems work in conjunction with humans, where the AI provides real-time insights, suggestions, and supports interactive decisions.	Real-time bidding in Digital Marketing, Augmented Reality (AR), Real-time Analytics	Collaborative Filtering, Collaborative Robots (Cobots)
Advisory	AI systems that provide suggestions or advice, yet final decision-making remains entirely with humans. They are primarily used to gain insights from data.	Data Analysis, Business Intelligence, Market Research	Unsupervised Learning, Anomaly Detection
Dependent	AI systems that are entirely dependent on human input for decision-making. They rely on humans for learning and improving their performance.	Supervised Learning in ML, Quality Control Systems, Sentiment Analysis	Active Learning, Feature Extraction

ETHICAL FRAMEWORKS HAVE BEGUN TO EMERGE

OECD AI principles (May 2019)

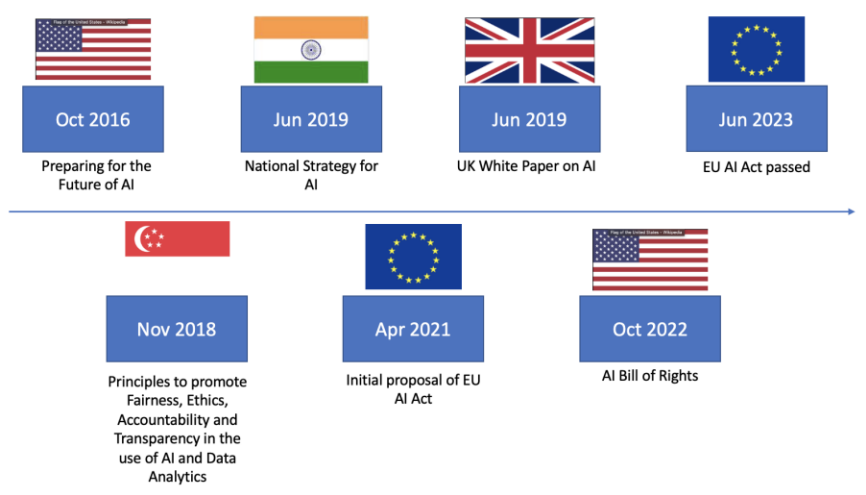


Behavioural Science "FORGOOD" (Lades and Delaney, 2022)

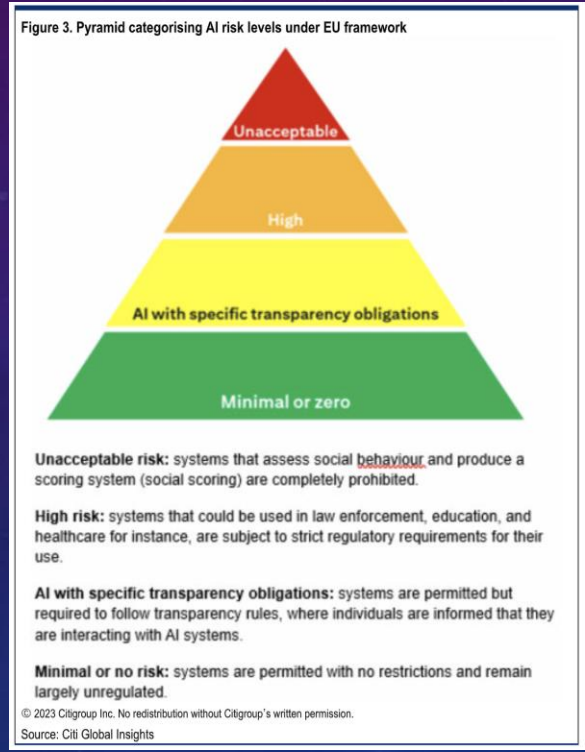


REGULATION

AI Regulation Timeline



Rules-based approach



Principles-based approach

Safety and robustness	US, UK
Protection against discrimination/ Fairness	US, UK, Singapore
Privacy	US
Transparency and explainability	US, UK, Singapore
Human alternatives, testing, and redress	US, UK
Accountability and governance	UK, Singapore

MACHINE BEHAVIOR – ASSESSING PRESENCE OF 'COGNITIVE BIAS' IN AI

Understand and uncover the patterns of 'decision-making in Artificial Intelligence, especially GenAI and chatbots to assess the extent these 'machines' exhibit patterns of thinking similar to humans

RATIONALES

Training Process / Data: Presence of patterns in training data without discernment.

Human Reinforcement Learning: Potential for bias to creep in with human input in model refinement

Assessment via:

- Replication of well-studied vignettes
- Novel experiments based on well-studied vignettes

Research is:

- Primarily focused on discrete choice experiments
- Limited research on assessing rationales for machine behaviour
- Primarily assess via qualitative methods with some using quantitative