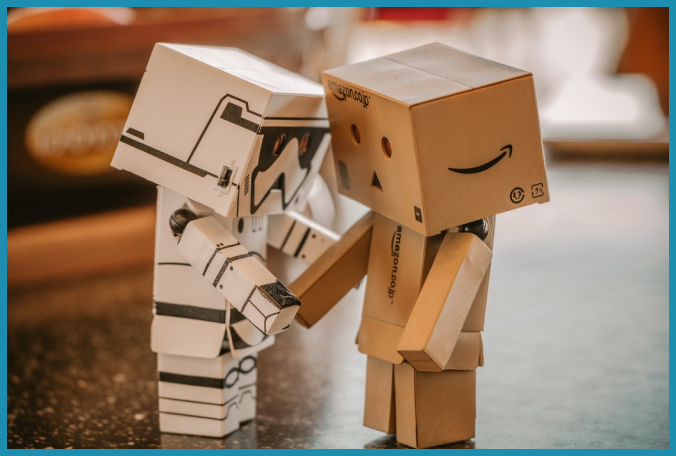
KU LEUVEN

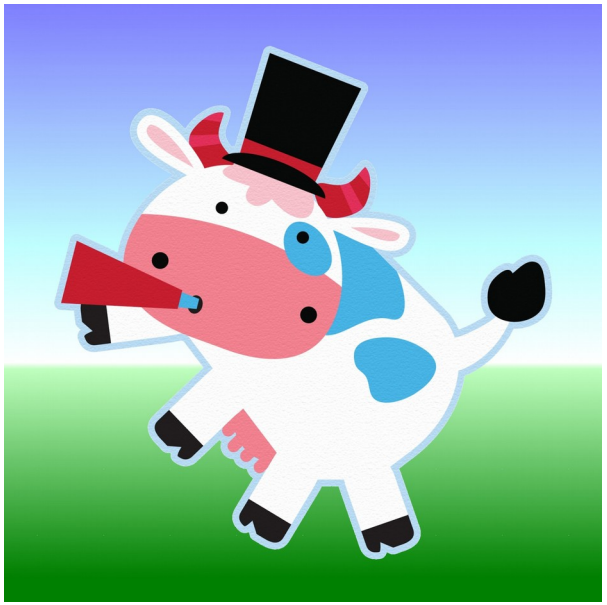
On engineering AI agents for privacy



Rafa Gálvez & Seda Gürses

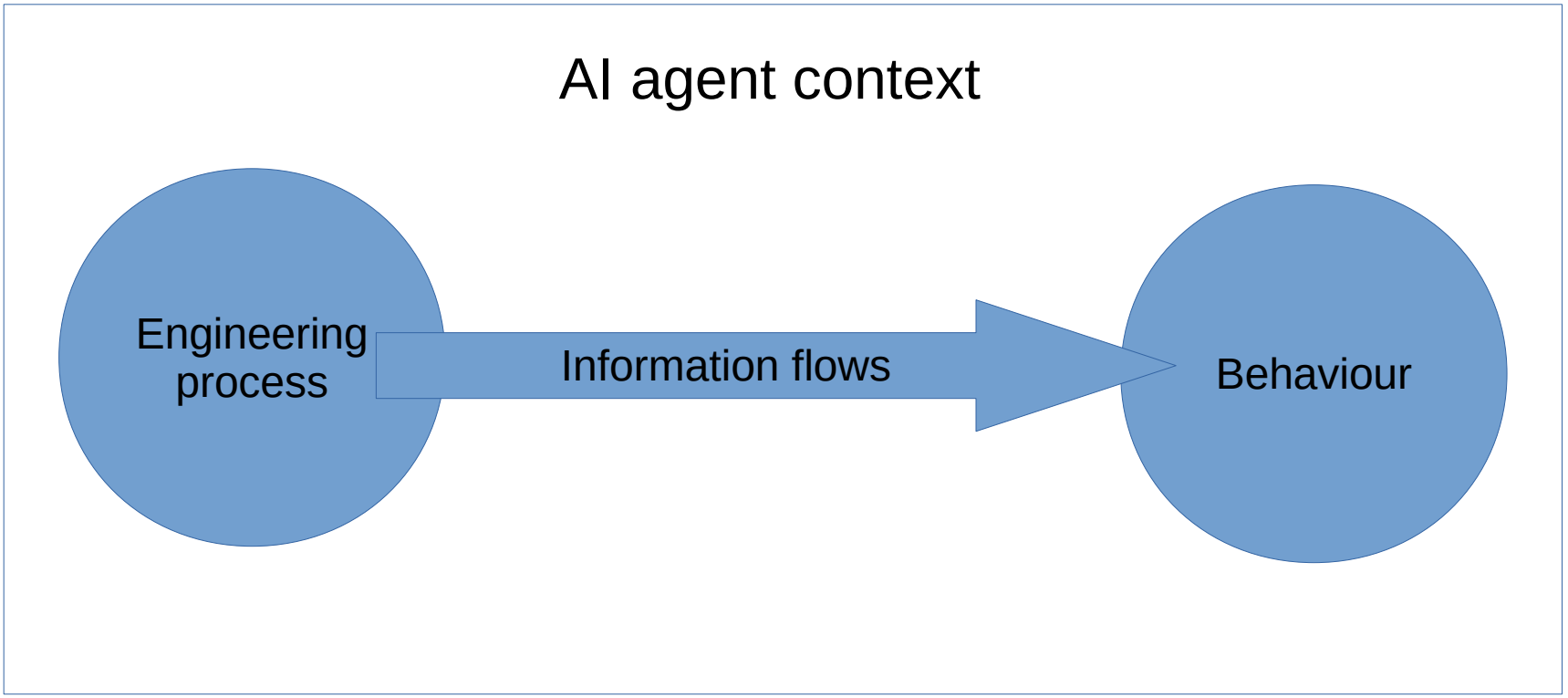
The motivation

Alexa, tell my child a story!



Not any story is OK, though

The problem


AI agent context
Engineering process
Information flows
Behaviour

The problem

	Engineering process		Behaviour
	Program	Model	Deployment
Information subject	Algorithm	Target population	Algorithm, Target population, users
Possible senders	Engineer, AI agent	Engineer, sample of population	Engineer, population sample, users
Possible recipients	Users	AI agent, Engineer	AI agent, users, engineer
Information types	Architecture, design choices, objective function	Stories, profile of users...	Architecture, design choices, objective function, stories, profiles of users...
Transmission principle	Public?	In confidence	Entitled by recipient, in confidence

Differences between types of information flows in the engineering process and the ones to enable behaviour

Why is this important?

Space stories for boys?

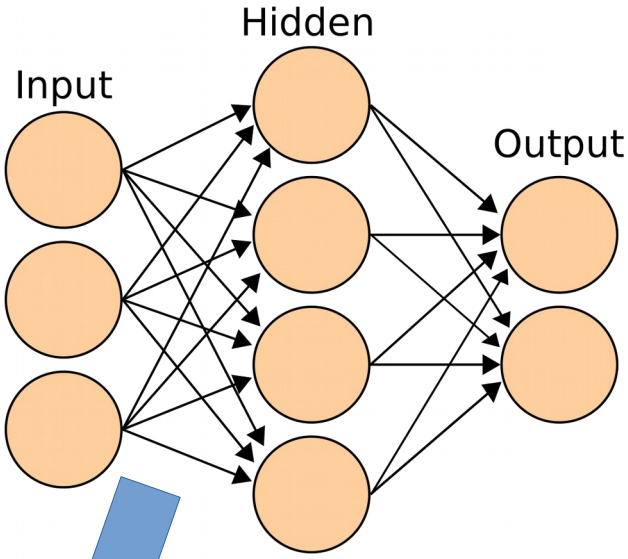
Princess stories for girls?

Recommendations very dependent on information flows: not all may be appropriate

Engineering vs Behaviour

A program...

performs tasks



A model

An AI agent performs intelligent tasks

Why do we use CI?

Engineers think about social norms

Engineers can also think about informational norms

Why do we use CI?

	Engineering process		Behaviour
	Program	Model	Deployment
Information subject	Algorithm	Target population	Algorithm, Target population, users
Possible senders	Engineer, AI agent	Engineer, sample of population	Engineer, population sample, users
Possible recipients	Users	AI agent, Engineer	AI agent, users, engineer
Information types	Architecture, design choices, objective function	Stories, profile of users...	Architecture, design choices, objective function, stories, profiles of users...
Transmission principle	Public?	In confidence	Entitled by recipient, in confidence

The design of an agent can expose questions about what are appropriate information flows

How do we use CI?

Machine Learning methodology, Géron'17

1. Look at the big picture.
2. Get the data.
3. Discover and visualize the data to gain insights.
4. Prepare the data for Machine Learning algorithms.
5. Select a model and train it.
6. Fine-tune your model.
7. Present your solution.
8. Launch, monitor, and maintain your system.

Feature selection

Better performance with feature *gender*

Is it appropriate to use *gender* to recommend stories?

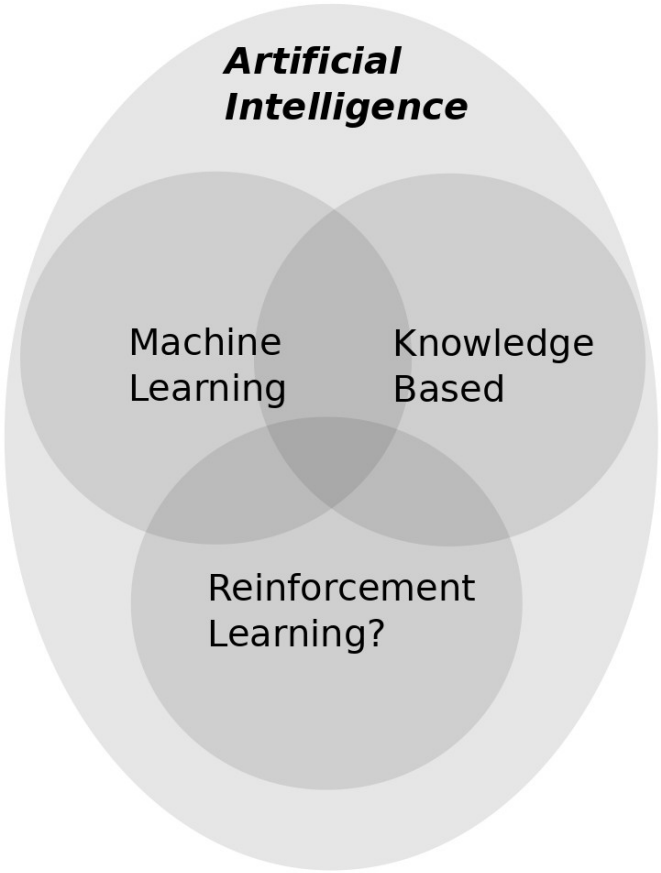
Current progress and results

1. Problem definition	1. Identify the task
2. Get the data	2. Assemble the relevant knowledge
3. Discover and visualize the data to gain insights	3. Decide on a vocabulary of predicates, functions, and constants
4. Prepare the data for machine learning algorithms	4. Encode general knowledge about the domain
5. Select, train and evaluate a model	5. Encode a description of the specific problem instance
6. Fine-tune the model	6. Pose queries to the inference procedure and get answers
7. Launch, monitor, and maintain the system	7. Debug the knowledge base

Current Machine Learning methodology

Current Knowledge Base methodology

Challenges encountered: AI



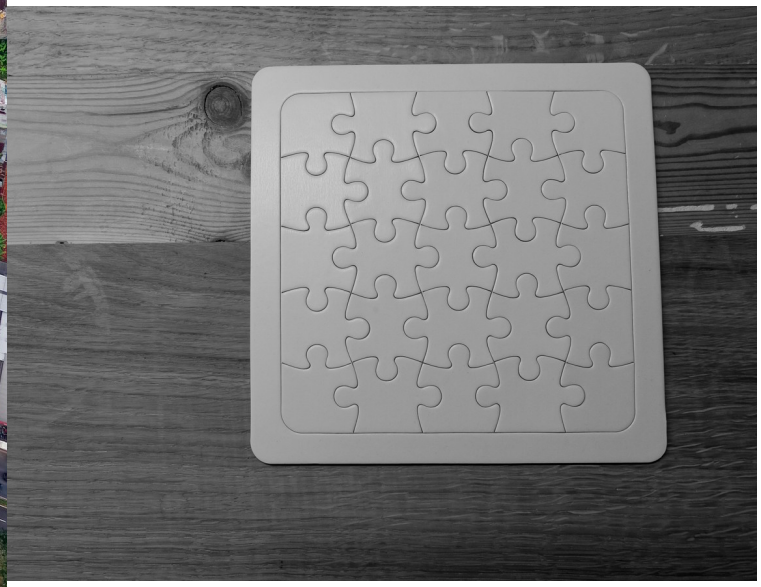
- What is an AI agent?
- How is Reinforcement Learning encompassing the whole AI field, including Knowledge Based and Machine Learning agents?
- Are there unique characteristics from RL that have an impact on the Contextual Integrity analysis?

Challenges encountered: CI

- **Single context**: are the engineering process and the behavior of the agent are in a single social context?
- **Appropriateness of flows in the case of inference**: is it inappropriate to use a proxy feature to get access to an inappropriate feature?
- **Composition of contexts**: what are the informational norms that govern mixed contexts?

Future work

Unified AI engineering methodology



Relate steps to the corresponding contexts