LEGAL INTENTION AND AUTONOMOUS DECISION-MAKING SYSTEMS

DAN HUNTER
ADM+S

dan.hunter@qut.edu.au

OVERVIEW

Overview

01

Introduction

In which our hero briefly reviews the rise of "AI" and examines the different models of XAI, concluding reluctantly that we are not going to be saved, not even by Geoff Hinton

02

Why does Intention matter?

A moment of legal instruction, brought to you by Jessica Fletcher and 264 episodes of "Murder, She Wrote"

03

Dominant Legal Models of ADMS Responsibility

A study of analogy (one of the greatest tricks of humankind) and its unfortunate limitations in law, by reference to the nature of slaves and companies

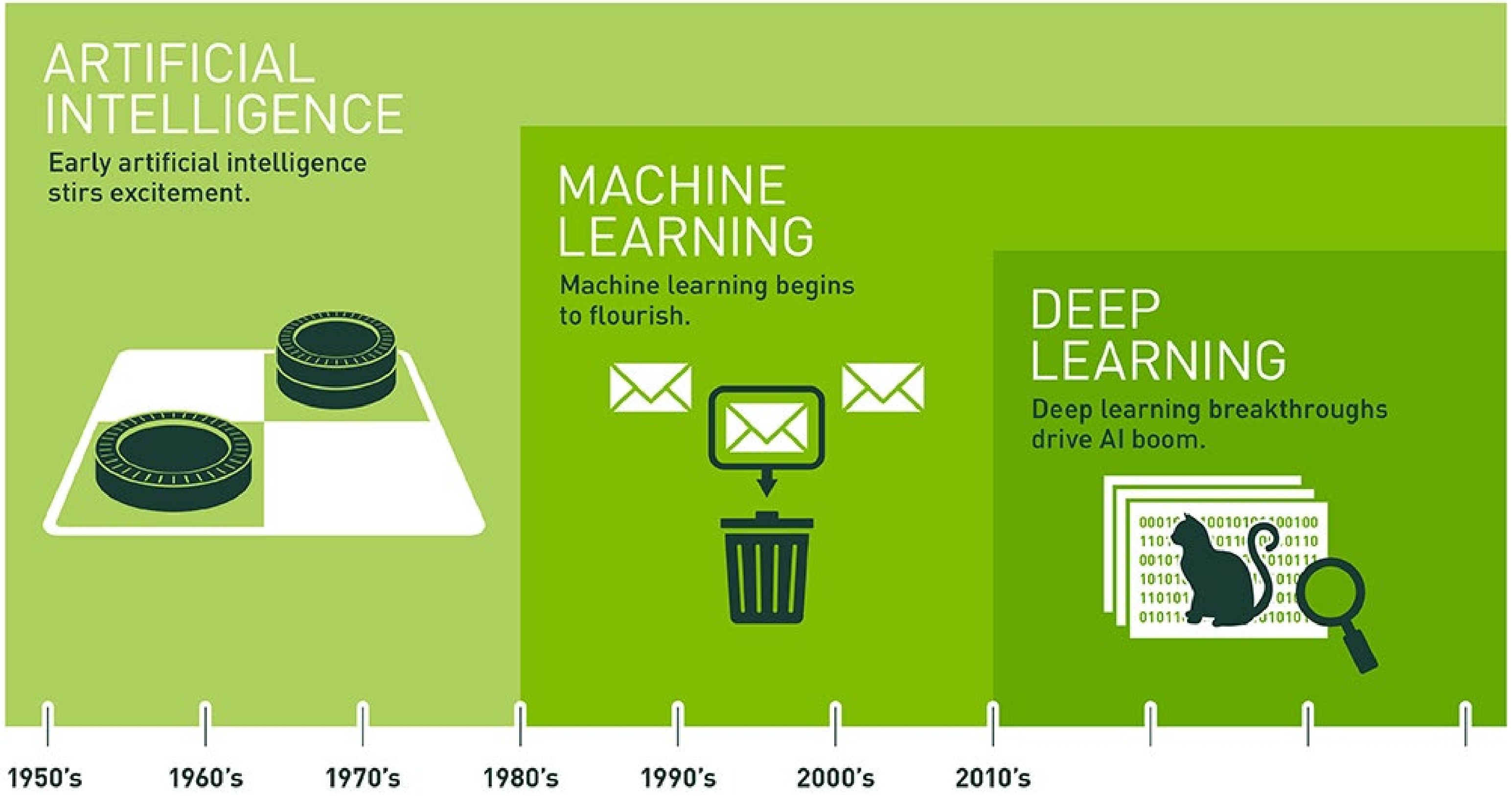
04

Legal Theories of Legal Intention

An all-too-brief examination of legal philosophy, delivered by someone who cares all-too-much about legal philosophy, delivered to people who care all-too-little

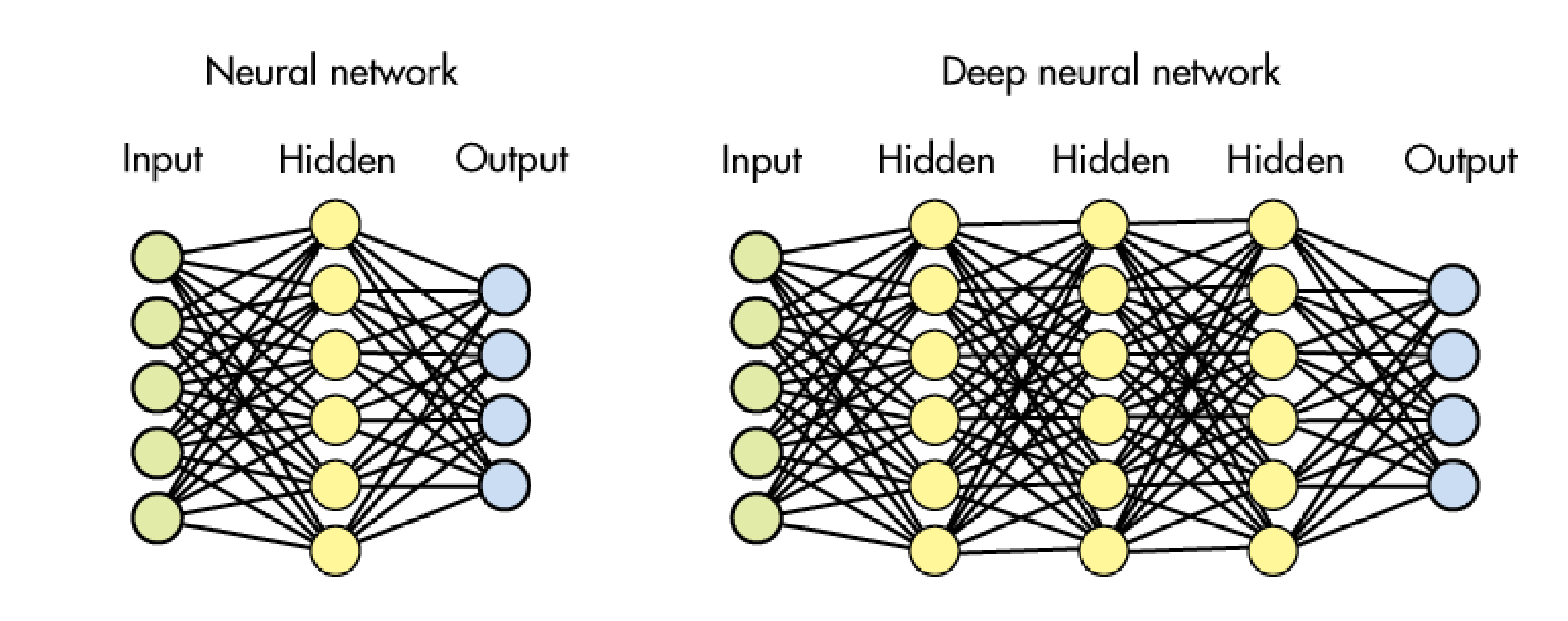
INTRODUCTION

Development of AI


ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Machine learning



The Black Box Problem, Interpretability and Explanatory AI

01

Model Dependent XAI

- Feature selection and component analysis - Štrumbelj and Kononenko (2014)
- NN rule extraction – Neurorule, TREPAN, Nefclass
- Layerwise Relevance Propagation – Bach et al (2015)

02

Model Agnostic XAI

- LIME (Local Interpretable Model-agnostic Explanations – Ribiero et al (2016)
- GIRP (Global Interpretation via Recursive Partitioning) - Yang et al (2018)

03

Counterfactuals

- Chen et al (2018)– minimal perterbations to change output
- Doshi-Velez & Kim (2017)

04

Neural Additive Models

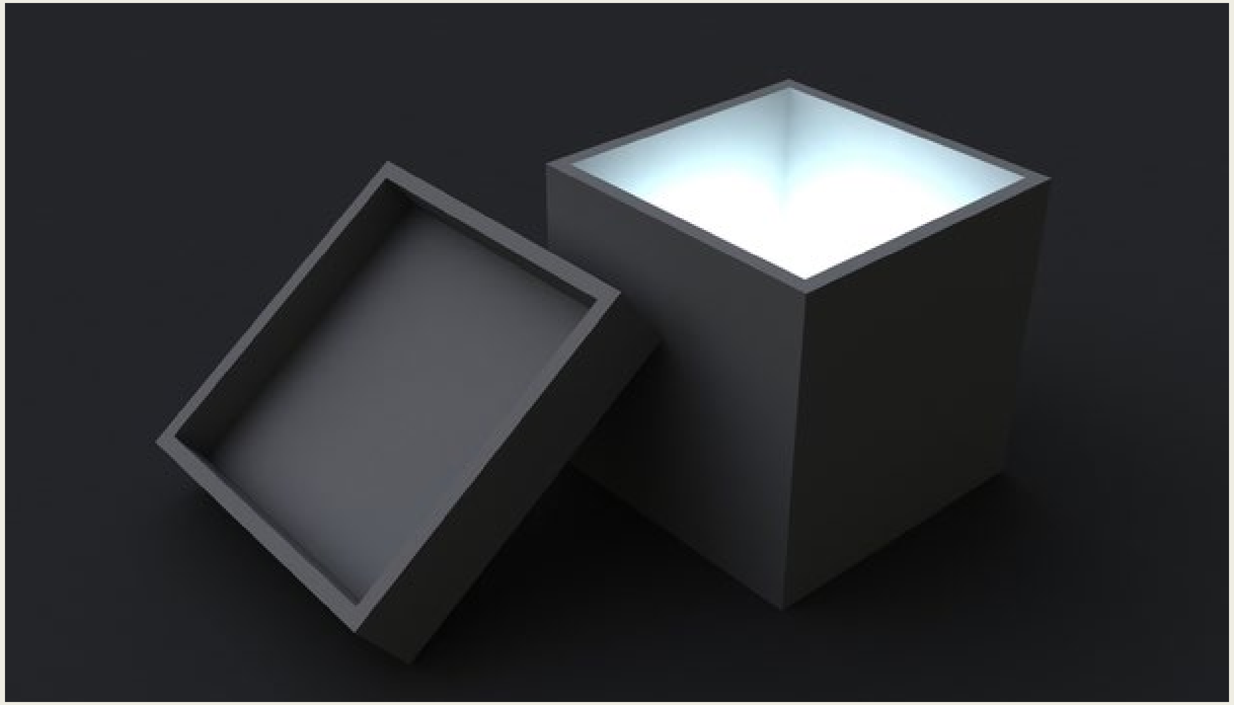
- Subset of Generalized Additive Models to create multiple structural models of the DNN - Agarwal et al (including Caruana & Hinton!) (2020)

The Black Box Problem, Interpretability and Explanatory AI (2)

At the end of the day we are still left with three problems:

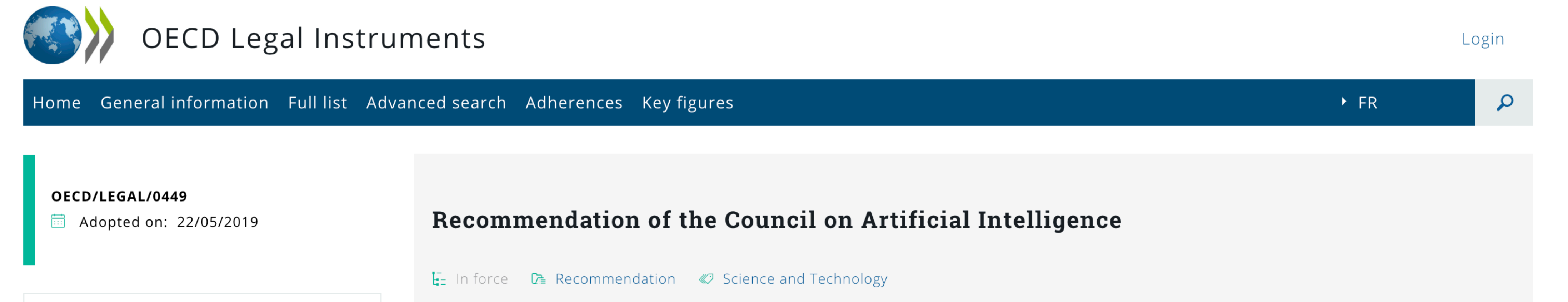
1. It’s a black box
2. The ADMS can act but it can’t think
3. No matter how sophisticated the post hoc XAI is, we are always going to be struggle to find intention inside the box.

   Aren’t we?



WHY DOES INTENTION MATTER?

Why do we care about ADMSs and Legal Intention? (1)



OECD Legal Instruments

OECD/LEGAL/0449

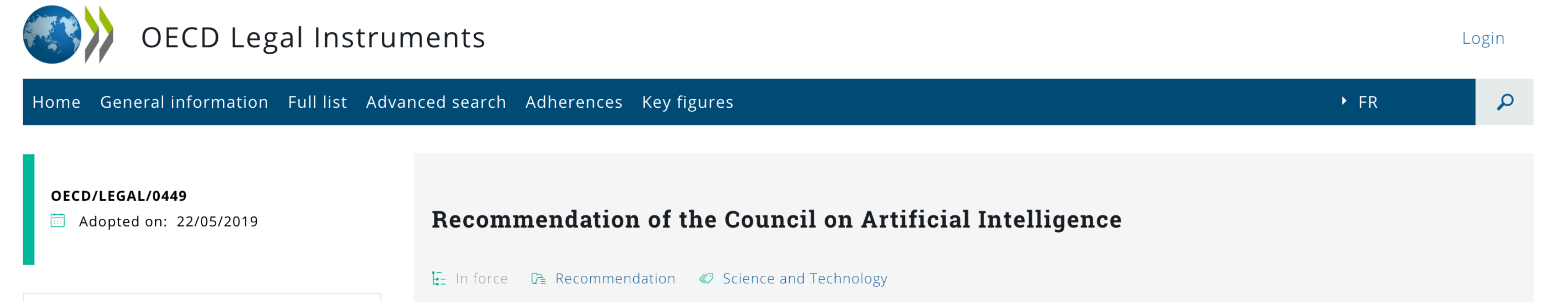
Adopted on: 22/05/2019

Recommendation of the Council on Artificial Intelligence

In force Recommendation Science and Technology

Section 1: Principles for responsible stewardship of trustworthy AI

… **RECOMMENDS** that Members and non-Members adhering to this Recommendation (hereafter the "Adherents") promote and implement the following principles for responsible stewardship of trustworthy AI, which are relevant to all stakeholders.



OECD Legal Instruments

OECD/LEGAL/0449

Adopted on: 22/05/2019

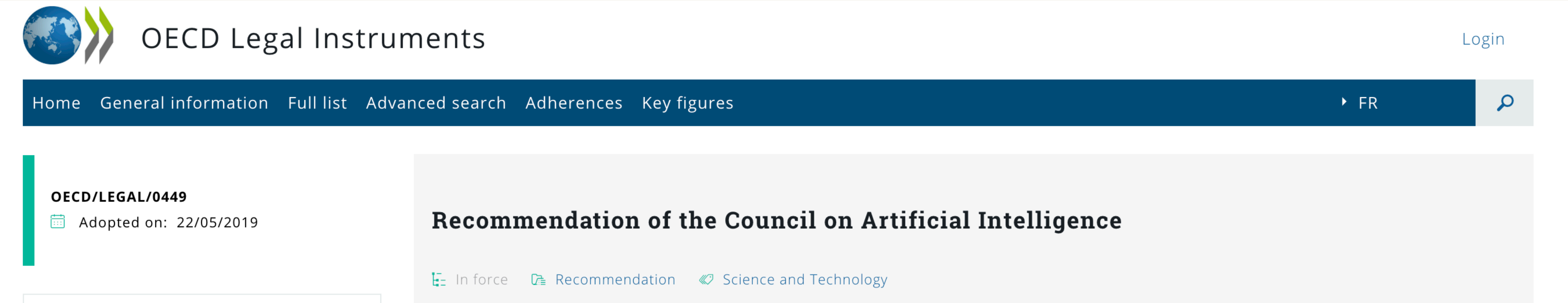
Recommendation of the Council on Artificial Intelligence

In force · Recommendation · Science and Technology

1.2. Human-centered values and fairness

a) AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights.

b) To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.



OECD/LEGAL/0449

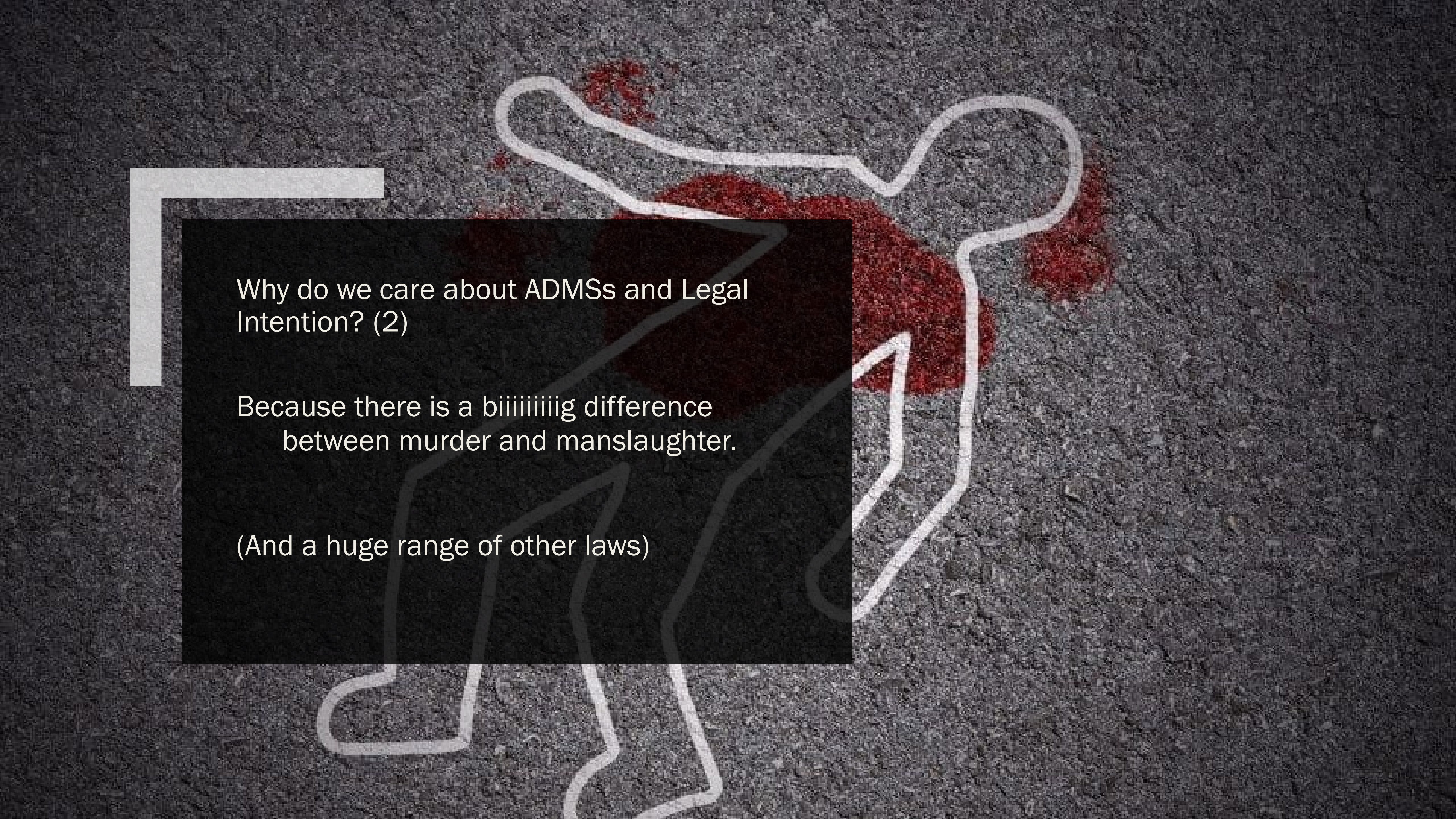
Adopted on: 22/05/2019

Recommendation of the Council on Artificial Intelligence

In force · Recommendation · Science and Technology

1.5. Accountability

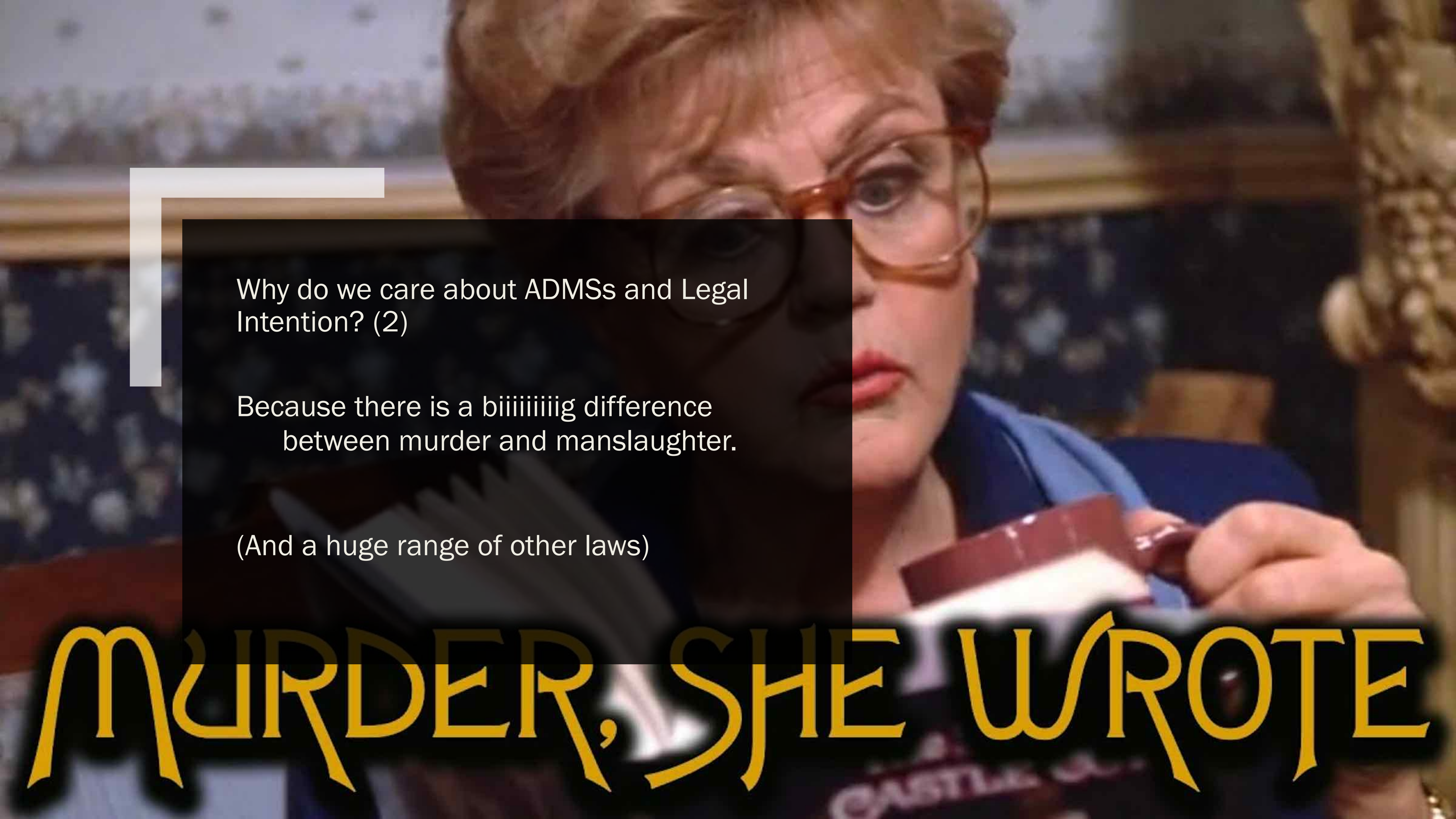
AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.



Why do we care about ADMSs and Legal Intention? (2)

Because there is a biiiiiiiiig difference between murder and manslaughter.

(And a huge range of other laws)

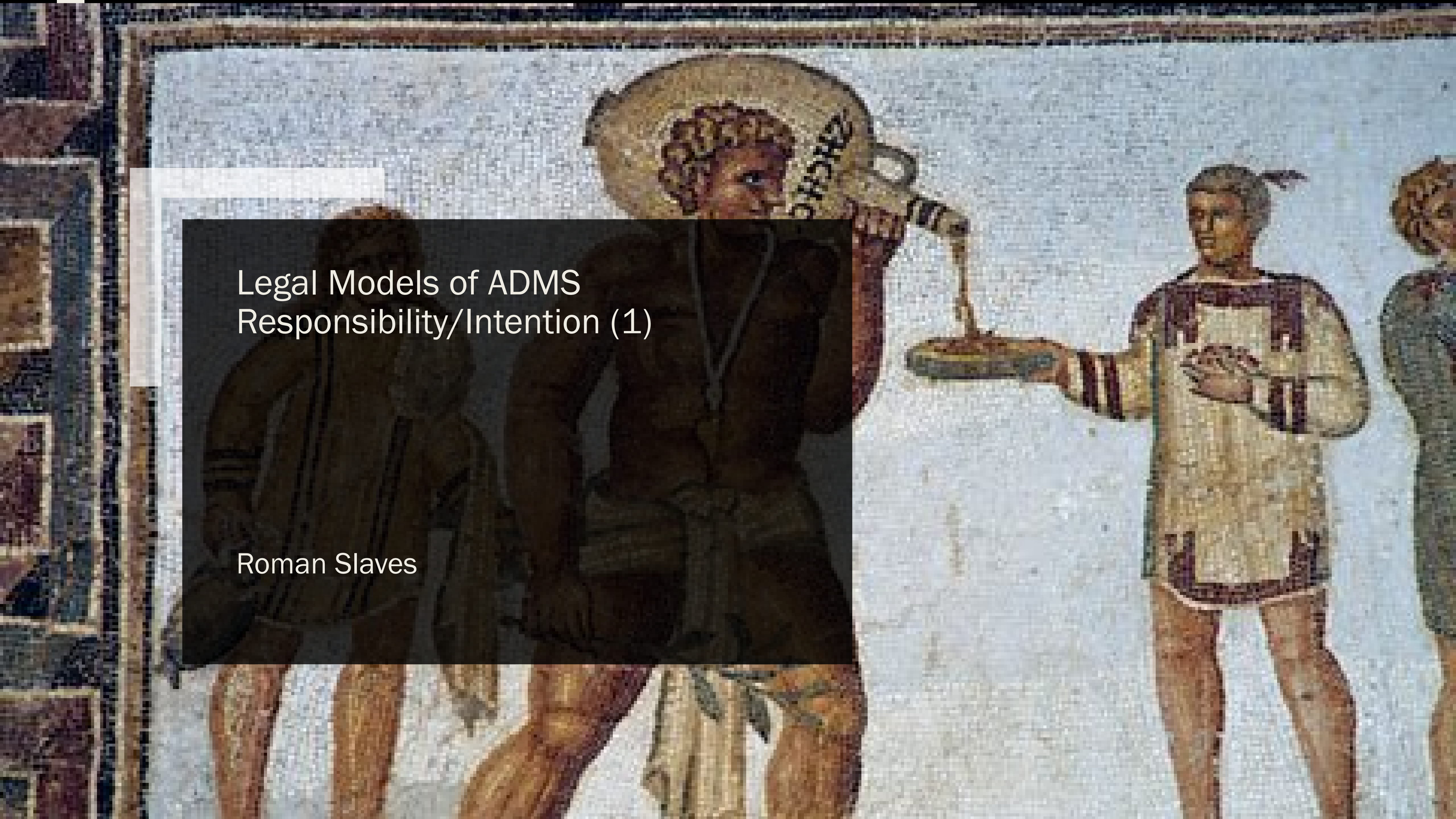


Why do we care about ADMSs and Legal Intention? (2)

Because there is a biiiiiiiig difference between murder and manslaughter.

(And a huge range of other laws)

DOMINANT LEGAL MODELS OF ADMS RESPONSIBILITY

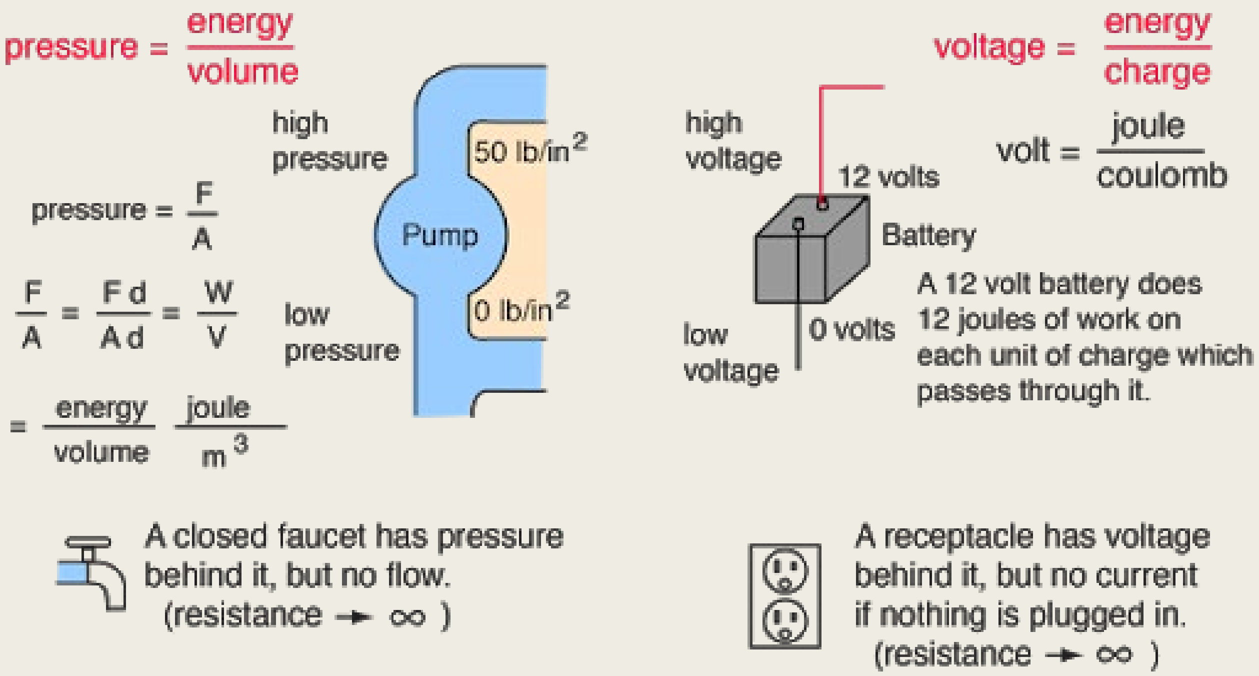


Legal Models of ADMS Responsibility/Intention (1)

Roman Slaves

Legal Models of ADMS Responsibility/Intention (2)

19th - 21st Century Companies


pressure = energy / volume
high pressure
50 lb/in2
pressure = F / A
Pump
F / A = F d / A d = W / V
low pressure
0 lb/in2
= energy / volume joule / m3
A closed faucet has pressure behind it, but no flow.
(resistance → ∞)
voltage = energy / charge
high voltage
12 volts
volt = joule / coulomb
Battery
A 12 volt battery does 12 joules of work on each unit of charge which passes through it.
low voltage
0 volts
A receptacle has voltage behind it, but no current if nothing is plugged in.
(resistance → ∞)

The Problems with Slaves and Firms

- It’s an (imperfect) analogy
- Not a principled understanding of intentionality

Solution?

- Principle-based legal theories of intentionality applied to ADMSs

LEGAL THEORIES OF LEGAL INTENTION

Theories of Legal Intention – Cliff's Notes edition

01

Formalist Theories

Focused on the consciousness of the agent, and their capacity to control their actions. Accordingly, there is a line of reasoning connecting 'what a person knows and intends, and what they do'.

Examples in criminal law (*actus reus* and *mens rea*)

02

Qualitative Theories

Moral qualitative analysis: ie, the agent's intentional actions are reflections of the agent's moral character.

Duff: An agent acts with intention if they: (i) want X to occur; (ii) believe their actions will or might realise X; and (iii) act on the basis of that desire and belief.

Natural law version of 1.

03

Functionalist Theories

The function of intention in law is in redistributing power across social spheres

Examples in contract and torts.

Cane: focus on the relationship between the agent's actions, the perceived seriousness and the consequences prescribed to it, operating within the broader social context

04

Non-Intent Theories

Regulatory decision to cut through the niceties and install some kind of non-intent recourse/liability model

E.g. dangerous driving, DUI, insurance or tribunal models (workplace safety) etc.

Theories of Legal Intention – Applicability to ADMSs

01

Formalist Theories

(Ignore 'person' for the moment)

Does the machine know and intend what it does?

Well, XAI models say yes…?

02

Qualitative Theories

Hard to attribute a moral state to a machine.

Basically a category error(?) and so an Aristotelean analysis seems impossible.

03

Functionalist Theories

This structural approach has some traction.

Focus on the power relations between the actors as indicative of the 'objective' intention of the machine

04

Non-Intent Theories

Most likely application because the alternatives are too hard

However, it lacks subtlety/nuance/context/precision and will fail.

> ‘It is quite clear from the reported cases of that, if a man in fact adopts a manner of driving which the jury think is dangerous to other road users in all the circumstances, then on the issue of guilt, it matters not whether he was deliberately reckless, careless, momentarily inattentive or doing his incompetent best.’
>
> *-- R v Webb (per Williams J.)*

This may be the answer

(Even though it stems from a functionalist/non-intent model)

CONCLUSION

What do we make of this?

1. Geoff Hinton is a god.

2. XAI connected with principle-based intention theories will get us some way down the track

3. The normal legal approach of analogizing (to slaves or companies) is probably a dead end, but one that will be hugely influential.

Dan Hunter
dan.hunter@qut.edu.au

END