HRI Reading Group

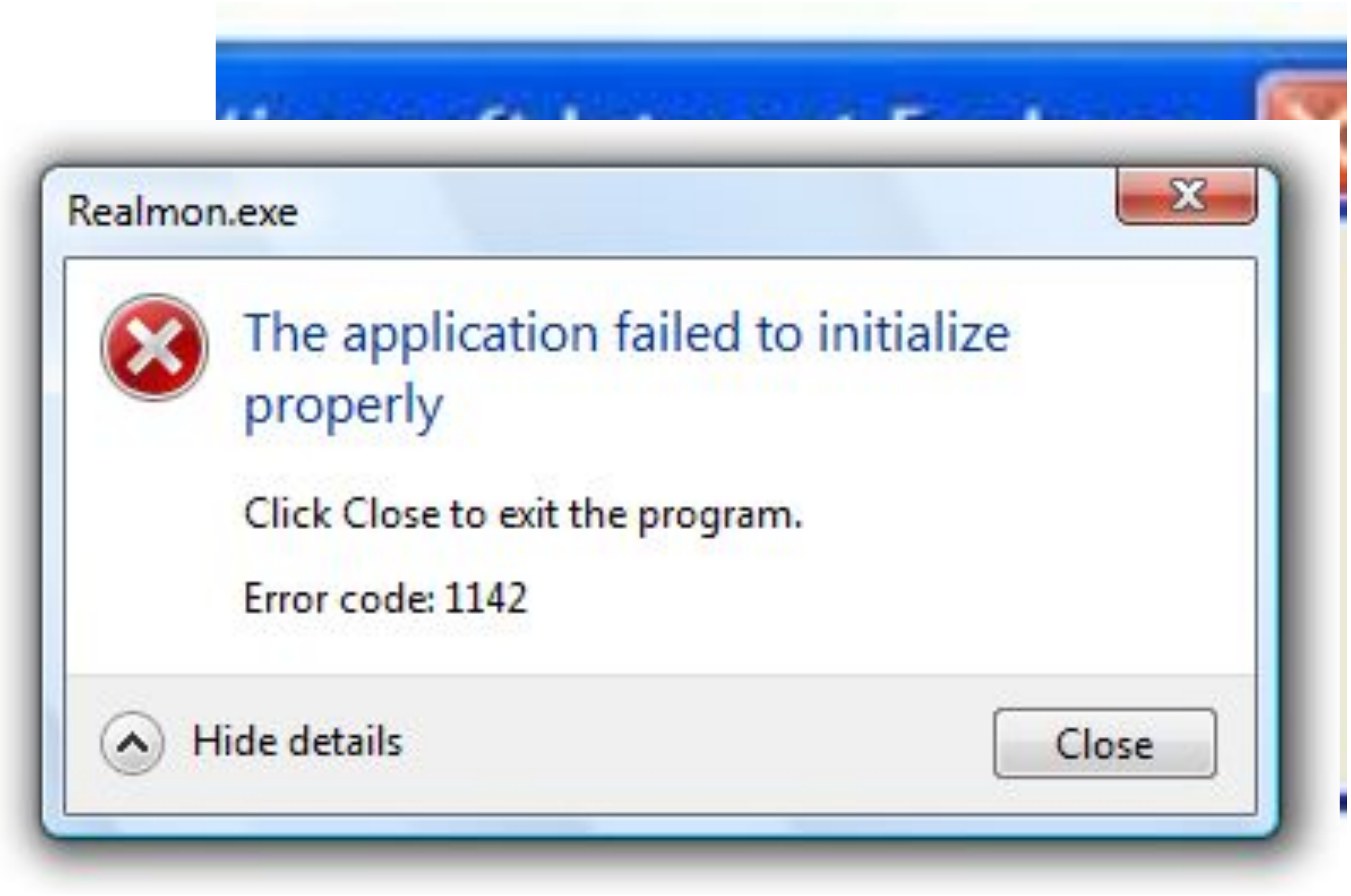
@ Instituto Superior Técnico
Spring 2019

Meeting #13 (June 7, 2019)

Paper

Rosenfeld, A. & Richardson, **"Explainability in human–agent systems"**
A. Auton Agent Multi-Agent Syst (2019).
https://doi.org/10.1007/s10458-019-09408-y

ERROR



Realmon.exe

The application failed to initialize properly

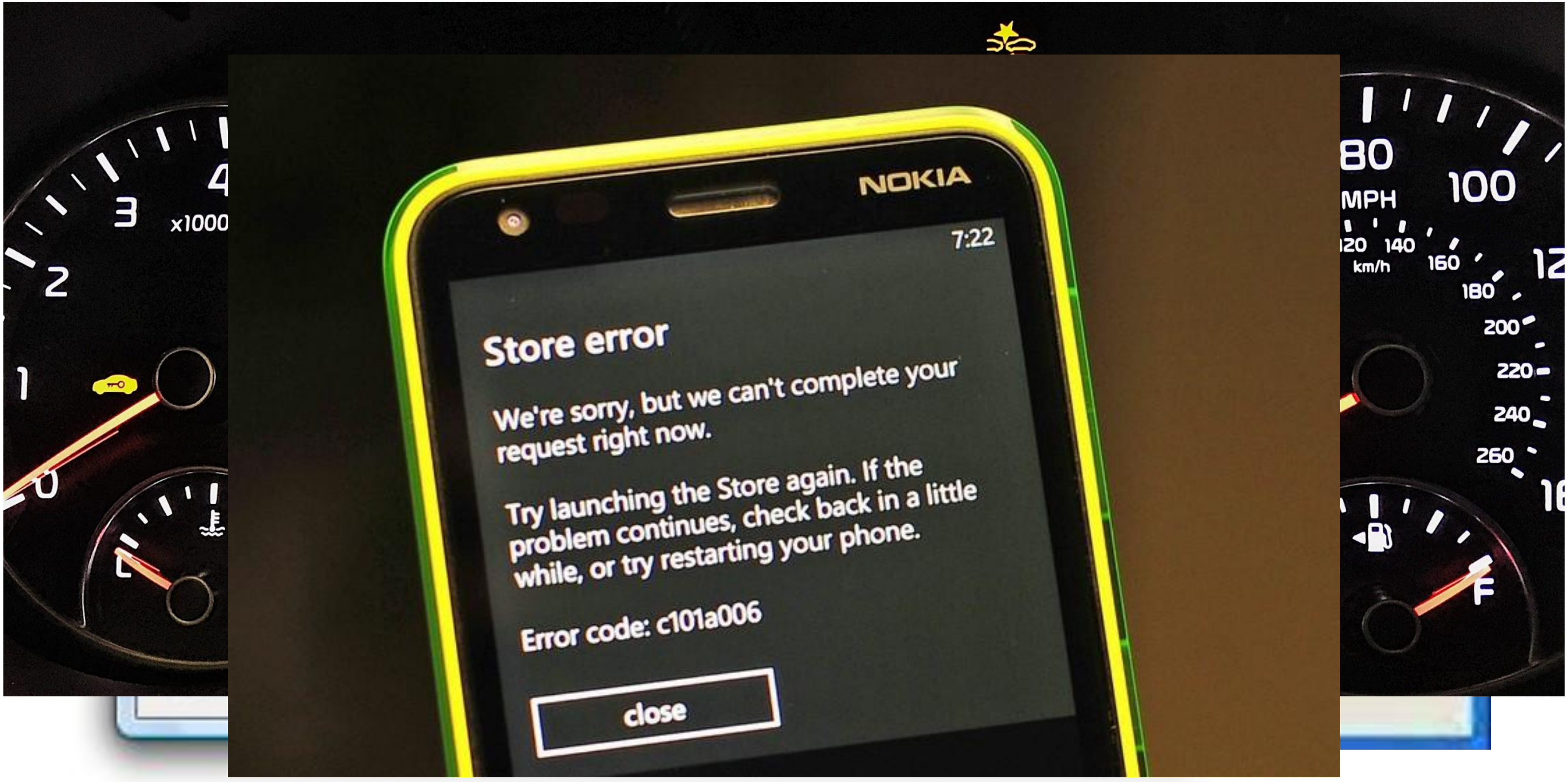
Click Close to exit the program.

Error code: 1142

Uh Oh

You're screwed.

x1000 rpm
1
2
3
4
5
6
7
8
0
C
H
System Check
22mi
BRAKE
EPB
ABS
OFF
MPH
km/h
0
20
40
60
80
100
20
40
60
80
100
120
140
160
180
200
220
240
260
E
F



Store error

We're sorry, but we can't complete your request right now.

Try launching the Store again. If the problem continues, check back in a little while, or try restarting your phone.

Error code: c101a006

Have you used explainability (or related terms) in your work? How?

Explainability and transparency as synonyms - ability of the robot to communicate internal states.

Communicating what the system can/cannot do (e.g., displays an Error message) and what the system is doing (as a scenario, e.g., “the robot cannot read”) are part of explainability? Debatable.

In the context of Machine Learning Algorithm transparency and explainability have a different meaning. Explainable Machine Learning is different from Explainable Planning.

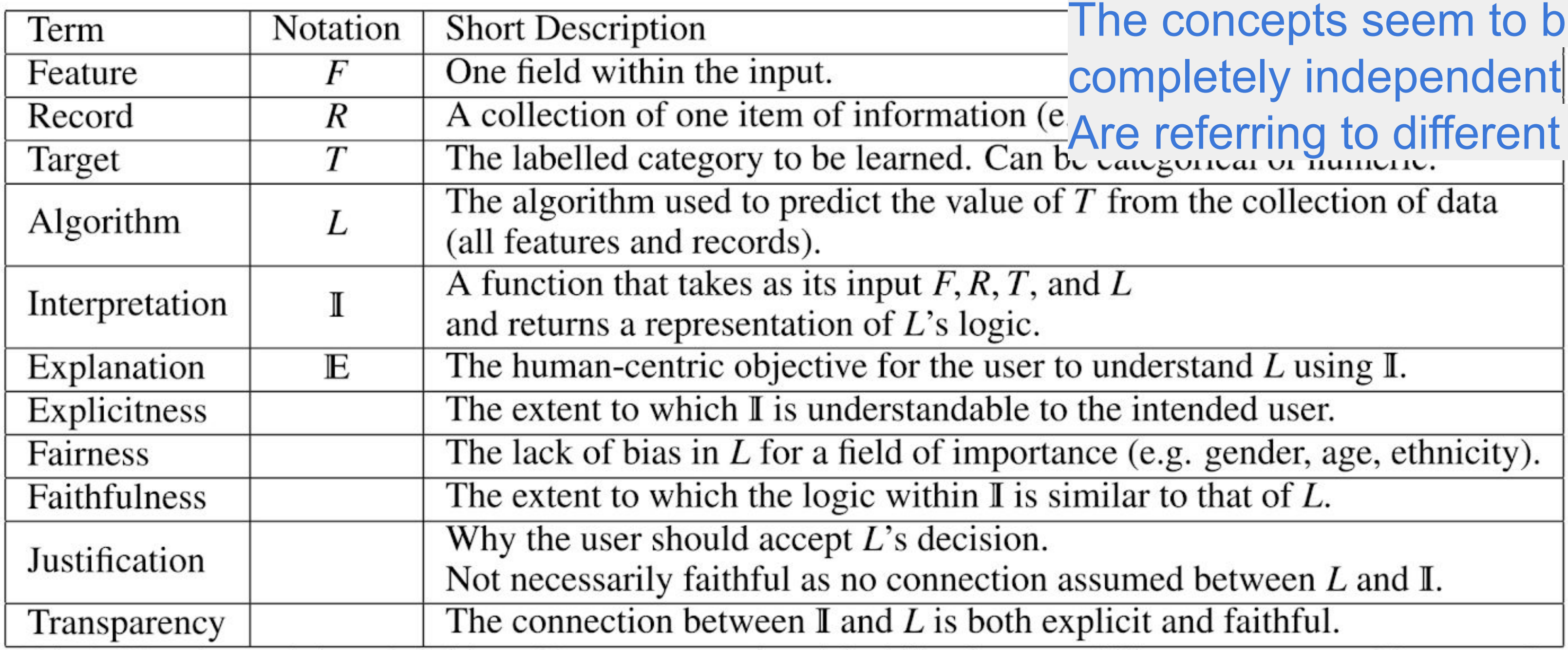
Although explainability can be viewed in relation to specific fields/methods (e.g,m ML algorithms, planning, etc), there are common aspects of explainability (such as human-system communication).

Have you used explainability (or related terms) in your work? How?

- Deep Learning and privacy problem
- Transparency in HCI is totally different:
    - *Any change in a computing system, such as a new feature or new component, is **transparent** if the system after change adheres to previous external interface as much as possible while changing its internal behaviour [https://en.wikipedia.org/wiki/Transparency_(human%E2%80%93computer_interaction)]*

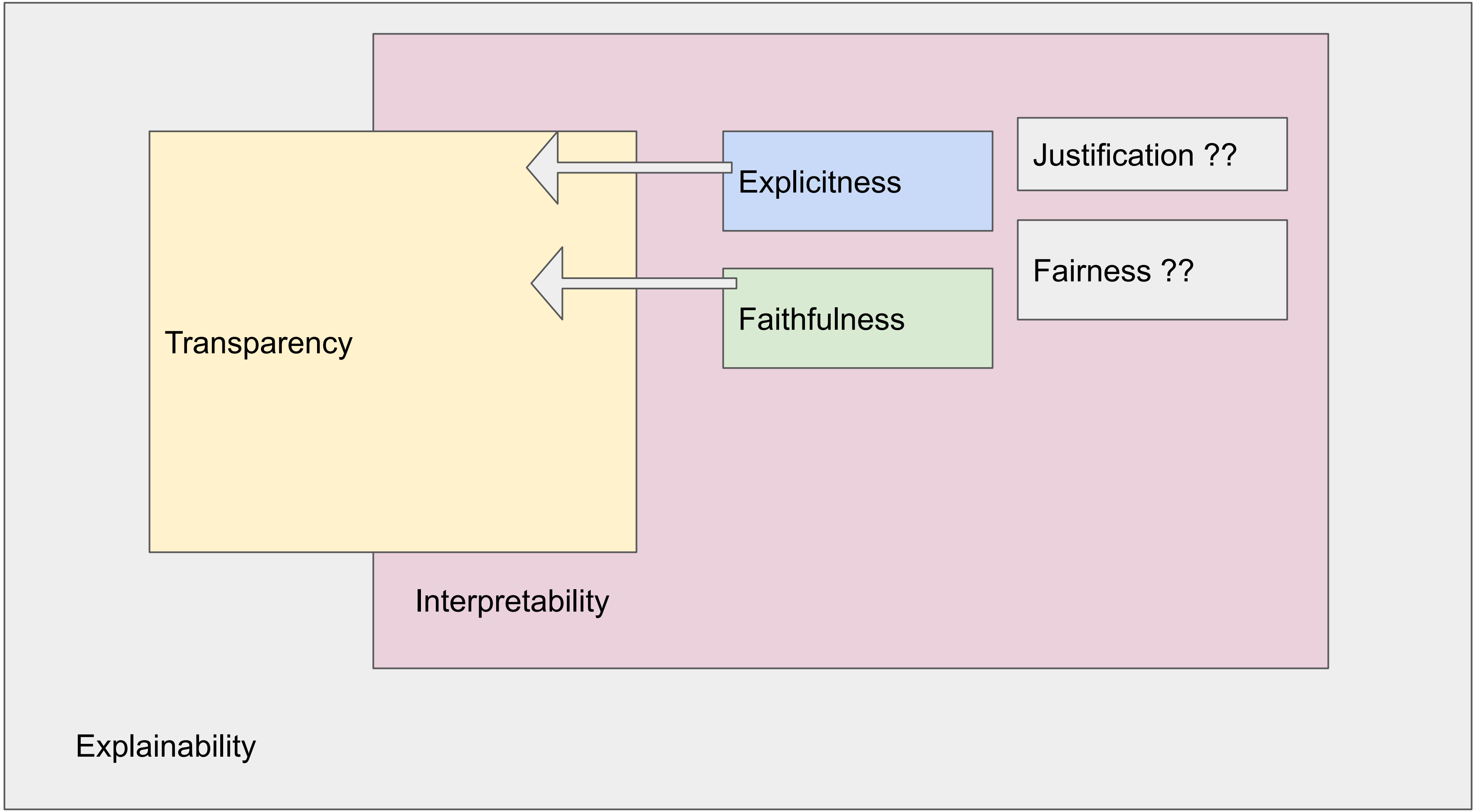
Explicitness is a scale
and

The concepts seem to be not completely independent
Are referring to different level



Term	Notation	Short Description
Feature	F	One field within the input.
Record	R	A collection of one item of information (e.
Target	T	The labelled category to be learned. Can be categorical or numeric.
Algorithm	L	The algorithm used to predict the value of T from the collection of data (all features and records).
Interpretation	$\mathbb{I}$	A function that takes as its input F, R, T, and L and returns a representation of L's logic.
Explanation	$\mathbb{E}$	The human-centric objective for the user to understand L using $\mathbb{I}$.
Explicitness		The extent to which $\mathbb{I}$ is understandable to the intended user.
Fairness		The lack of bias in L for a field of importance (e.g. gender, age, ethnicity).
Faithfulness		The extent to which the logic within $\mathbb{I}$ is similar to that of L.
Justification		Why the user should accept L's decision. Not necessarily faithful as no connection assumed between L and $\mathbb{I}$.
Transparency		The connection between $\mathbb{I}$ and L is both explicit and faithful.

Table 1: Notation and short definition of key concepts of explainability, interpretability, transparency, fairness, and explicitness in this paper. Concepts of features, records, targets and machine learning algorithms and explanations are also included as they define the key concepts.


Justification ??
Explicitness
Fairness ??
Faithfulness
Transparency
Interpretability
Explainability

“(...) it is important to consider new machine learning algorithms that include explainability as a consideration *within the learning algorithm*.”

“Several of these methods use an element of feature analysis as the basis for their transparency”

Can we consider explainability (and related terms) the first step to “conscious machines”?

Why is explainability important in a system?

For who can a system’s explainability be?

What explanation should be generated by the system?

White box vs black box algorithms: what is the role of transparency?

Why does accuracy decrease with more explainability?

[add Fig 3]

Can systems have more than one explanation?

Many of the visualization approaches, because they are designers of the algorithms are the ones designing their explainability, produce interpretations that are not easily understood by people without an expert-level understanding of the problem being solved, making them not very explicit. Who should be invited to design the explanations?

General person as the user.

Linguistics experts. Filmmakers. Journalists. Psychotherapists. Psychologists.

A communication between different experts need to occur, namely from different backgrounds.

How can explainability relate to HRI?
The *why, who, when, what* of explainability

Consider the following scenarios:

Critical vs beneficial system explanations

Examples of critical

Examples of benefitial

Transparency

“Decision model where the decision-making process can be directly understood without any additional information” (Guidoti et al., 2018)

Decision trees are transparent, but deep neural networks are not.

Does this mean transparency is tied to non-learning models?

HRI Reading Group

@ Instituto Superior Técnico

Meeting #14 (14 June 2019)
Invited Moderator - Samuel Gomes