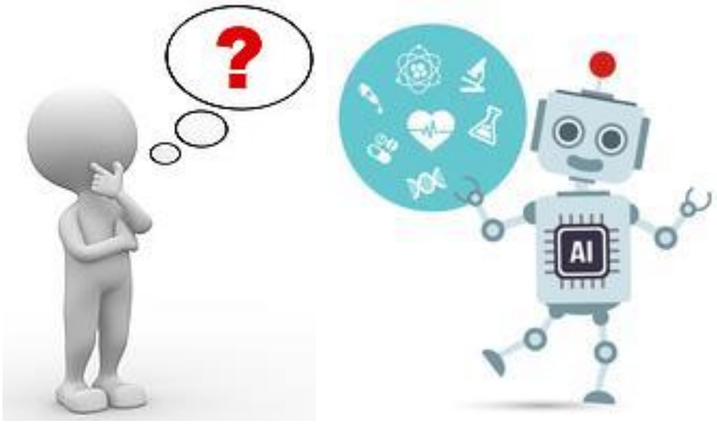
Risk and Exposure of XAI in Persuasion and Argumentation: The case of Manipulation

Rachele Carli[1], **Amro Najjar**[2], **Davide Calvaresi**[3]

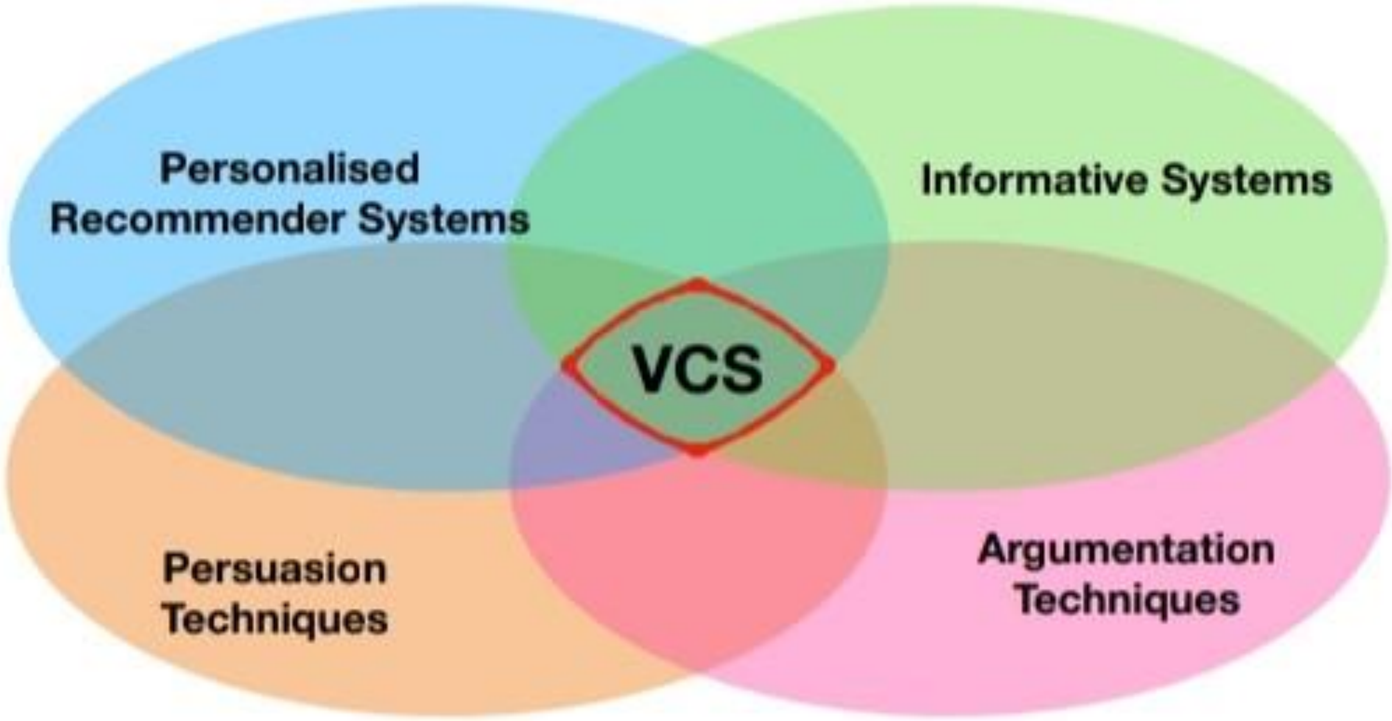
[1]Alma Mater Research Institute for Human-Centered AI, University of Bologna & ICR Group, University of Luxembourg;

[2]Luxembourg Institute of Science and Technology (LIST), Luxembourg;

[3]University of Applied Sciences Western Switzerland, Switzerland



Introduction



Explainabiliy

Transparency

Safety

Autonomy

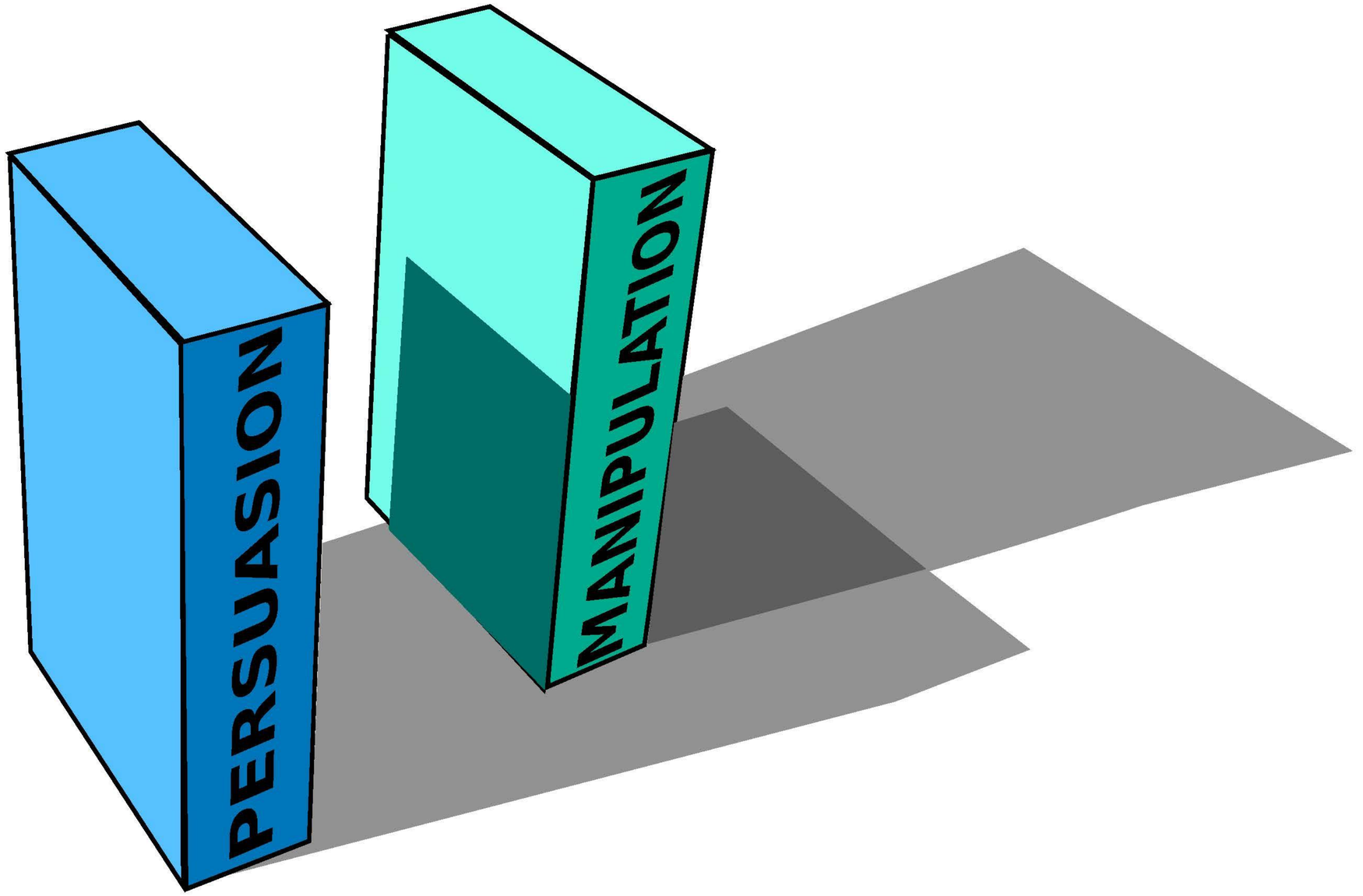
Open questions

Can AI systems be **actually** transparent for non-experts?

Can the **mere** fact of giving an explanation make the system safer?

Can the explanation make the user **really aware** of system's dynamics?

Persuasion vs Manipulation



Persuasion vs Manipulation

No clear definitions

No clearly identifiable causal link

No clear boundaries

Curbing **Self-governance**

Desiderata

Integrity
of the human
person
as a whole

Desiderata

Adaptivity

Granularity of explanations

Paternity of choice over autonomy

Order matters

From manipulation to manipulative techniques

Specificity vs General cases:

- Context-sensitive
- Goal-sensitive
- Precise interaction-sensitive

Desiderata

Adaptivity

Granularity of explanations

Paternity of choice over autonomy

Order matters

From manipulation to manipulative techniques

For each recommendation:

- Specific expl.
- Identifiable expl.
- No hidden implications
- Easier traceability

Desiderata

Adaptivity

Granularity of explanations

Paternity of choice over autonomy

Order matters

From manipulation to manipulative techniques

- Track decision-making process
- Conscious reconstruction of decision
- Coherence with users' goal

Desiderata

Adaptivity

Granularity of explanations

Paternity of choice over autonomy

Order matters

From manipulation to manipulative techniques

- Relevance
- Highest impact on safety
- FRIA criterion

Desiderata

Adaptivity

Granularity of explanations

Paternity of choice over autonomy

Order matters

From manipulation to manipulative techniques

- Correcting he dynamics to correct the effect
- No exploitation of vulnerabilities
- Maximisation of users' utility
- Respect for users' goal/will

Questions?

rachele.carli2@unibo.it

amro.najjar@list.lu

davide.calvaresi@hevs.ch