INTANGIBLE ASPECTS OF TRUSTING AUTONOMOUS SYSTEMS

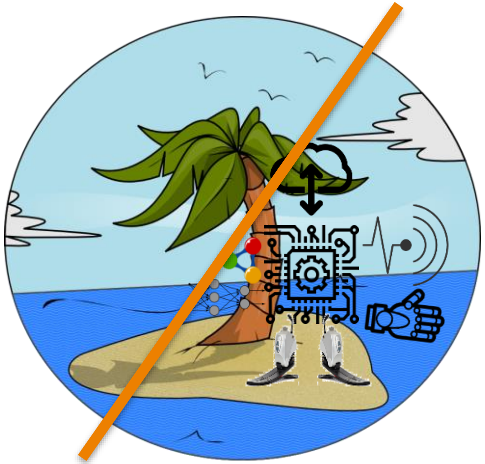
Mark Neerincx

TNO innovation for life

TU Delft
Delft University of Technology

INTRODUCTION

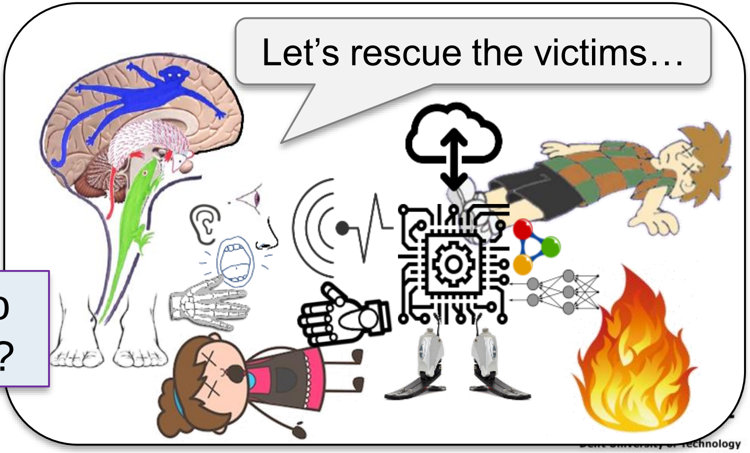
- No Autonomous System is an Island.
  - i.e, there is intelligence, interaction (incl. sensing & acting) and some kind of embodiment
  - let's call it robot, which -physically or virtually- embodies some kind of Artificial Intelligence and acts in a dynamic environment with other actors



- We aim at Human-Robot collaboration
  - i.e., responsible and effective hybrid teaming
  - in which both the robot and the human mutually adapt and learn over time

- Realizing that perception, cognition and behavior of humans and robots are fundamentally different

How do they relate to each other over time?



HUMAN-ROBOT RELATIONSHIPS…

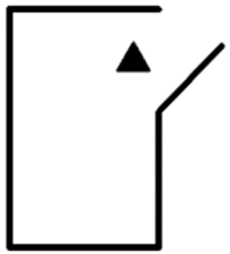
- Anecdotes of Explosive Ordnance Disposal (EOD) operators in Iraq and Afghanistan:
  - robots were assigned names and gendered identities
  - when a robot was damaged, its loss was grieved, sometimes accompanied by funeral-like rituals
  - when a robot had to be repaired, its operators requested to fix, instead of replace, its mechanical parts, to preserve the robot’s individual identity
  - in rare occasions, soldiers have endangered themselves to protect the robot from enemy assaults

AGENCY AND ATTACHMENT

- Humans show an instinctive tendency to attribute animacy and intentions even to entities that have little or no resemblance at all to animated or living creatures

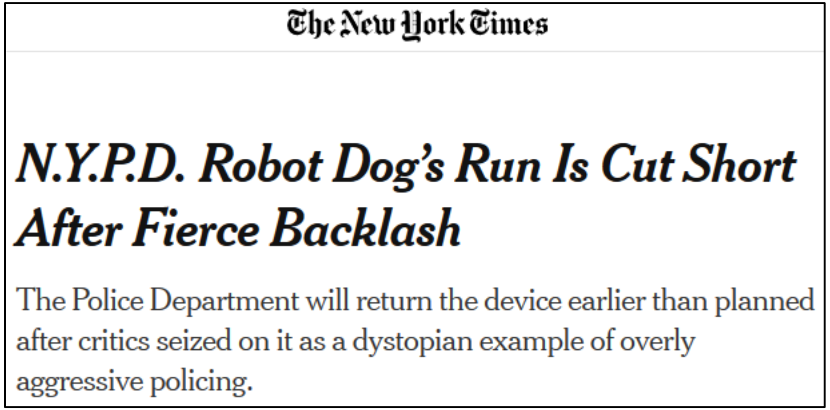
- Humans get attached to technology

- Three perspectives on human-robot relationships:
  - Human-Companion: robot as an electronic partner (ePartners)
  - Human-Human: robot as a human (humanoid/android robot, anthropomorphism)
  - Human-Animal: robot as an animal (zoomorphism)



Heider & Simmel experiment (1944).

HUMAN-ANIMAL TEAMWORK



The New York Times

N.Y.P.D. Robot Dog's Run Is Cut Short After Fierce Backlash

The Police Department will return the device earlier than planned after critics seized on it as a dystopian example of overly aggressive policing.

"We should have named it 'Lassie.'"

HUMAN-ANIMAL TEAMWORK PERSPECTIVE

Example shows:

- The animals or robots interact with humans in the team ***and*** with humans in their environment
- Transparency about animals or robots “role and goals” is crucial

Integration in team:

- History of incorporating animals into our work provides insights in how humans might deal with robots to augment team performance.
- Animals, with different perception, cognition and motor capabilities, have become powerful team members that enable us to work differently.

Initial trustworthiness:

- A person draws conclusions about the attributes, personality, capabilities, and level of intelligence of an animal, regardless of whether or not they are true characteristics, behaviors, or capabilities.

Trust development:

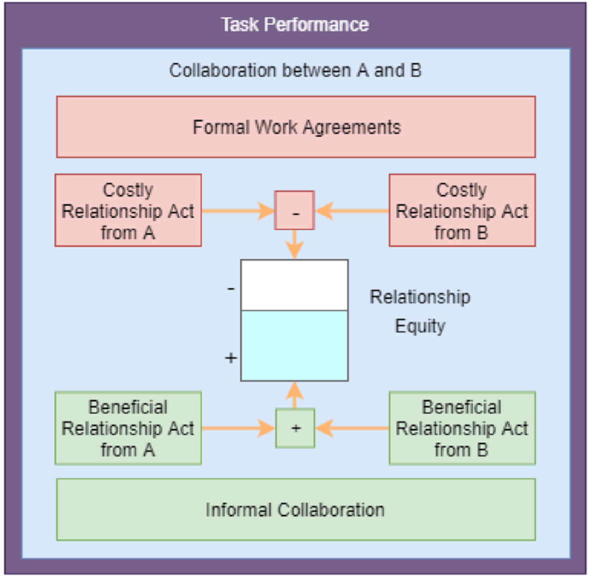
- Influenced by animal’s capabilities, situated predictability and predispositions of the person.
- Mutual trust is based on communication and respect, often a result of training.
- A successful partnership develops when humans interact with their animals regularly, enabling them to predict how that animal reacts to most situations
- The riskier the situation is, the more important human-animal trust becomes.

HUMAN-HUMAN TEAMWORK

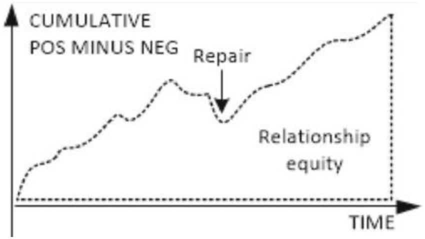
- Mutual trust is a fundamental property and predictor of high performing teams.

- Trust is a relational concept, i.e.,
  - “The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party”.
- Interpersonal relationship development is based on social exchange,
  - sharing and trading resources is a fundamental aspect of relationships, including intangible resources.

- Trust develops as a function of experience, i.e.,
  - trust depends on persistent, competent behavior of a party that pursues a desired goal



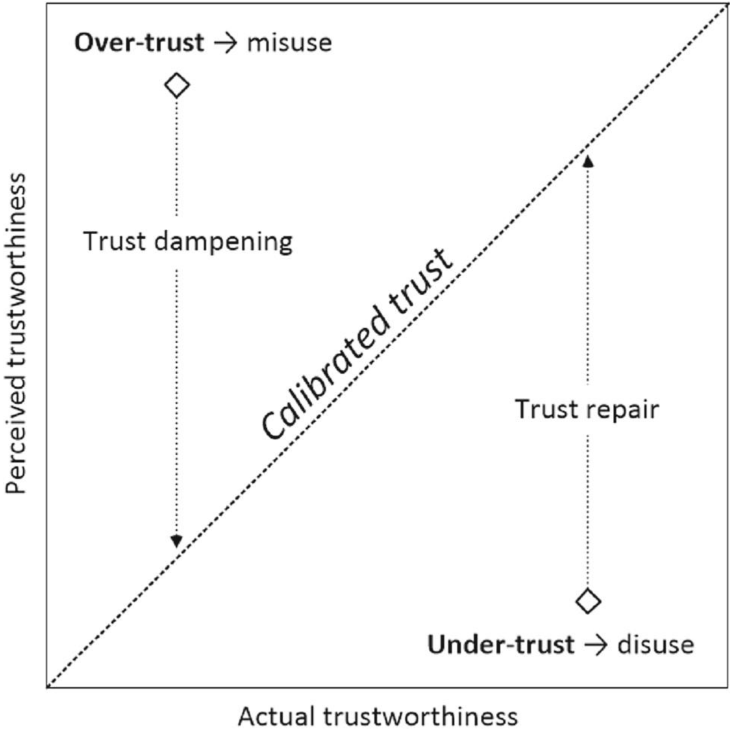
Task Performance
Collaboration between A and B
Formal Work Agreements
Costly Relationship Act from A
-
Costly Relationship Act from B
-
Relationship Equity
+
Beneficial Relationship Act from A
+
Beneficial Relationship Act from B
Informal Collaboration



HUMAN-EPARTNER TEAMWORK…

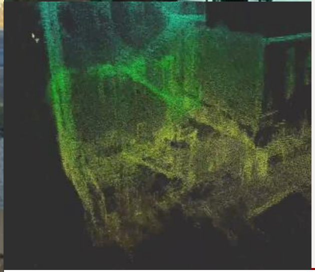
Calibration mechanisms:

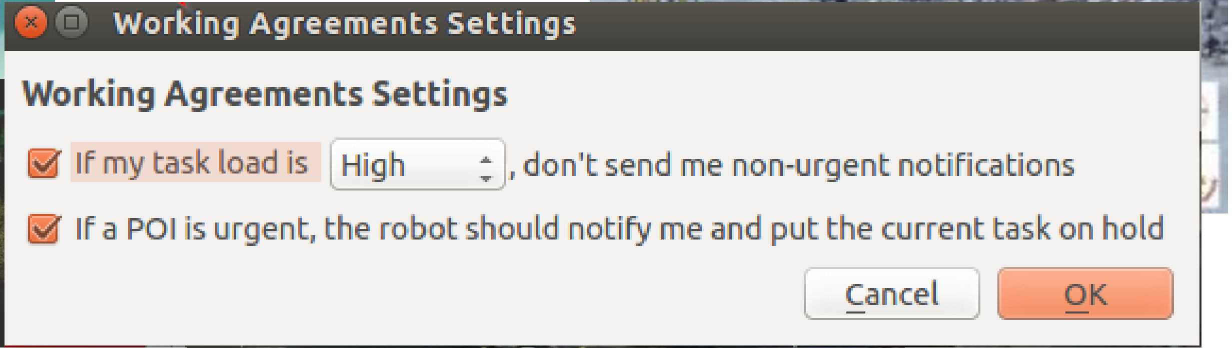
- Transparency discloses information about processes and states, improving interpretability
  - Interpretable confidence measure
- Explanation clarifies the relations between information entities, improving understandability
  - Contrastive explanation
- Sharing experiences supports learning and personalization
  - Cognitive-affective memory
- Work agreements support predictability
  - Commitment model



Trustworthiness is the extent to which an actor has the ability to execute relevant tasks, demonstrates integrity, and is benevolent towards fellow team members

WORK AGREEMENT EXAMPLE





Working Agreements Settings

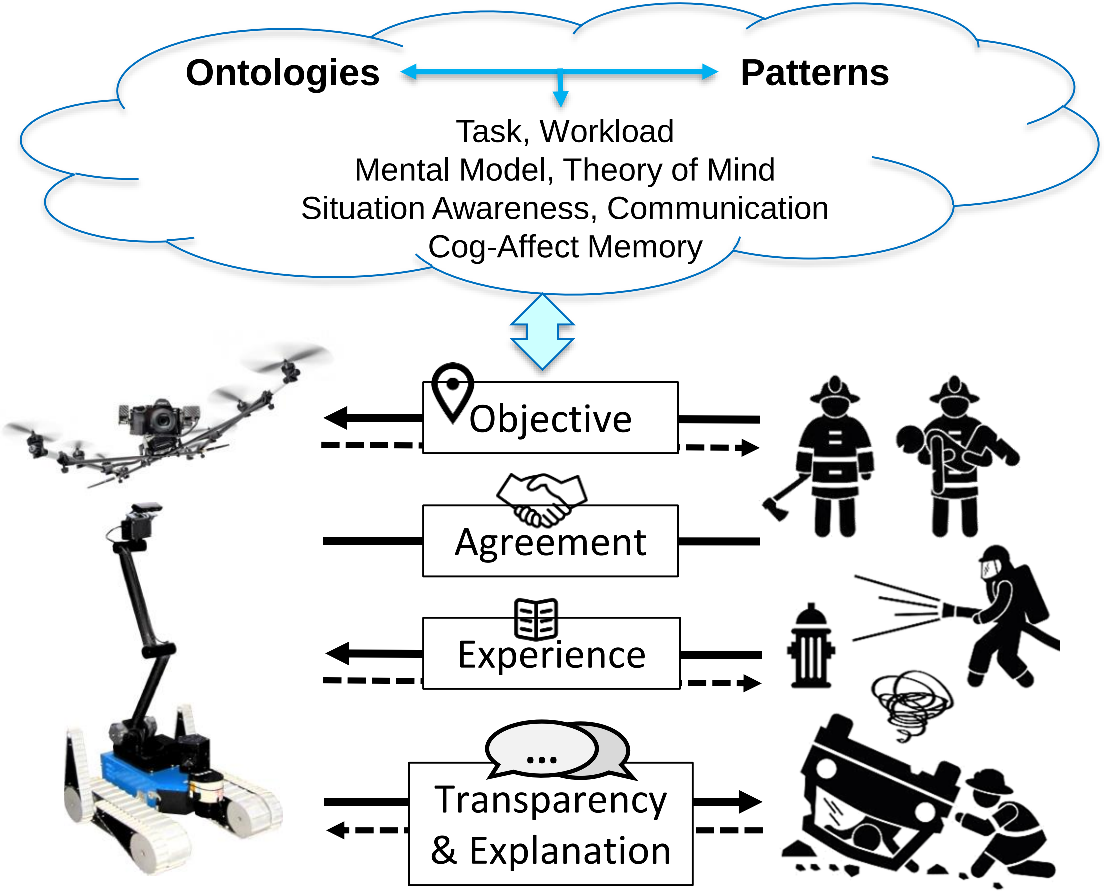
Working Agreements Settings

☑ If my task load is High, don't send me non-urgent notifications

☑ If a POI is urgent, the robot should notify me and put the current task on hold

Mark Neerincx

HUMAN-EPARTNER TEAMWORK



Example Projects

- **Space** (MECA)
- **Disaster Management** (NIFTi, TRADR & ASSISTANCE)
- **Railway**
- **Naval C4ISR**

Mark Neerincx

CONCLUSIONS

- Humans and robots are distinct, forming a new type of team (called hybrid teams), warranting new theorizing and modeling (particularly given the variation in robot roles, skills and embodiments).

- Robot integration in teams will not only bring about new human–robot relationships but will also change human-human relationships.

- Human-human and human-animal relationship development can inspire the design of human-robot relationships and trust calibration methods.

- Four trust calibration mechanisms advance hybrid teaming:
  - Transparency
  - Explanations
  - Experience sharing
  - Work agreements

KEY REFERENCES

Cappuccio, M. L., Galliott, J. C., & Sandoval, E. B. (2021). Saving Private Robot: Risks and Advantages of Anthropomorphism in Agent-Soldier Teams. *International Journal of Social Robotics*, 1-14.

Carpenter, J. (2013). *The Quiet Professional: An investigation of US military Explosive Ordnance Disposal personnel interactions with everyday field robots* (Doctoral dissertation, University of Washington).

Darling, K. (2021). *The New Breed: What Our History with Animals Reveals about Our Future with Robots*. Henry Holt and Co., New York.

De Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics*, *12*(2), 459-478.

Fox, J., & Gambino, A. (2021). Relationship Development with Humanoid Social Robots: Applying Interpersonal Theories to Human/Robot Interaction. *Cyberpsychology, Behavior, and Social Networking*. Vol. 24(5). DOI: 10.1089/cyber.2020.0181

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660.

Heider & Simmel (1944). An Experimental Study of Apparent Behavior". *American Journal of Psychology.* 57, 243–259

Johnson, M., & Vera, A. (2019). No AI is an island: the case for teaming intelligence. *AI Magazine*, *40*(1), 16-28.

Mioch, T., Peeters, M. M., & Neerincx, M. A. (2018). Improving adaptive human-robot cooperation through work agreements. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 1105-1110). IEEE.

Neerincx, M. A., van der Waa, J., Kaptein, F., & van Diggelen, J. (2018). Using perceptual and cognitive explanations for enhanced human-agent team performance. In *International Conference on Engineering Psychology and Cognitive Ergonomics* (pp. 204-214). Springer, Cham.

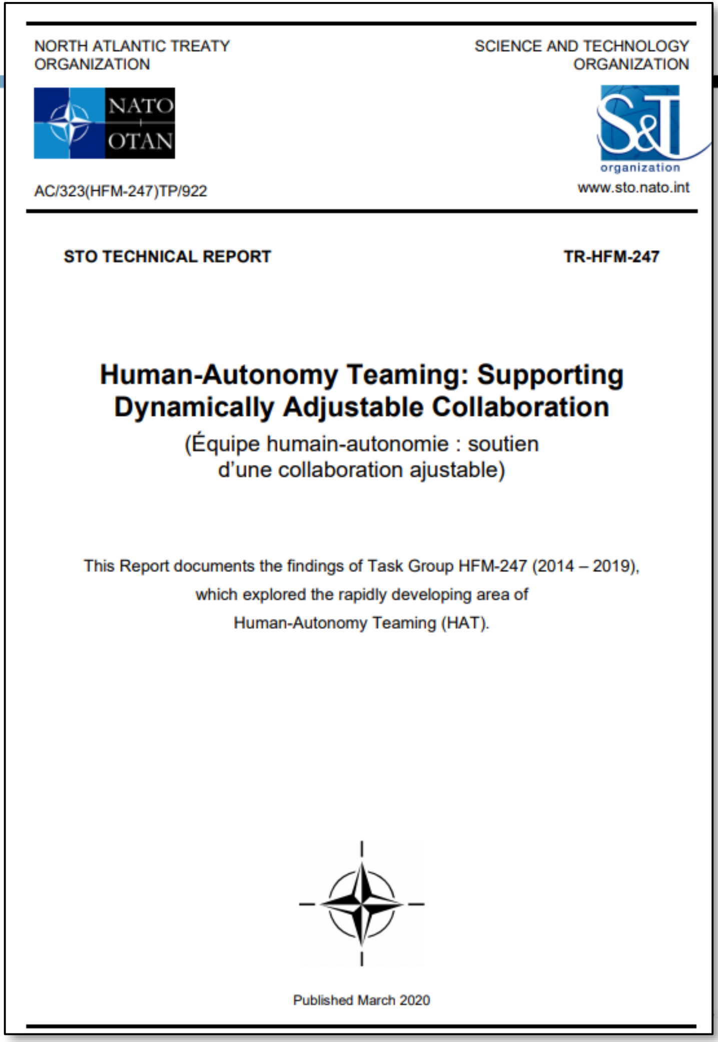
Phillips, E., Schaefer, K. E., Billings, D. R., Jentsch, F., & Hancock, P. A. (2016). Human-animal teams as an analog for future human-robot teams: Influencing design and fostering trust. *Journal of Human-Robot Interaction*, 5(1), 100-125.

Van der Waa, J., Schoonderwoerd, T., van Diggelen, J., & Neerincx, M. (2020). Interpretable confidence measures for decision support systems. *International Journal of Human-Computer Studies*, *144*, 102493.

HFM-247 HUMAN-AUTONOMY TEAMING

Recommendations for further research and development:

1. Meaningful human control: How to establish and maintain across all AI systems
2. Team design patterns for dynamic evolving behaviors
3. Continuous trust-calibration for proper reliance on automation
4. Scope enlargement to cover all relevant teaming structures and characteristics
5. Explainable AI in human-agent teamwork
6. Evolving hybrid intelligence by co-learning



NORTH ATLANTIC TREATY ORGANIZATION

SCIENCE AND TECHNOLOGY ORGANIZATION

AC/323(HFM-247)TP/922

www.sto.nato.int

STO TECHNICAL REPORT **TR-HFM-247**

Human-Autonomy Teaming: Supporting Dynamically Adjustable Collaboration

(Équipe humain-autonomie : soutien d'une collaboration ajustable)

This Report documents the findings of Task Group HFM-247 (2014 – 2019), which explored the rapidly developing area of Human-Autonomy Teaming (HAT).

Published March 2020

THANKS FOR YOUR ATTENTION

mark.neerincx@tno.nl