Applying Dempster Shafer Theory to Validate AAI Scales as Measures of Moral Agency in Robots

Nathanael Parra

MirrorLab Dr. Tom Williams

Purpose

This research aims to explore how humans perceive moral agency in robotic systems by examining specific operational characteristics—autonomy, adaptability, and interactivity (A, A, I). These traits are foundational to the framework proposed by Floridi and Sanders for assessing artificial agents.

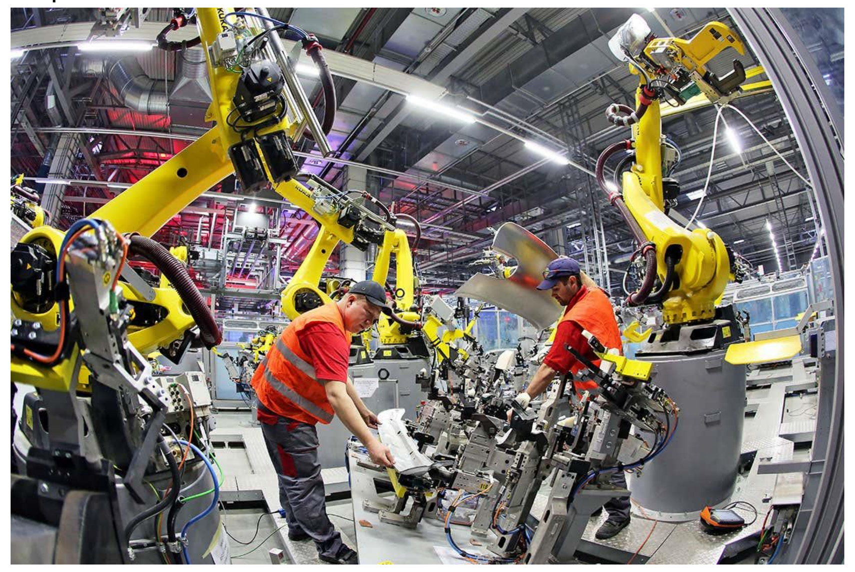
- **To validate Floridi and Sanders scales** of autonomy, adaptability and interactivity as measures of moral agency in robotic systems.
- **Use Dempster Shafer Theory** (DST) to generate novel statistics about human perception A,A,I measures given its ambiguous / uncertain nature.

By grounding this study in DST, we aim to move beyond traditional statistical approaches that require assumptions of clarity and binary responses. Instead, DST allows us to model uncertainty directly, capturing the nuanced and sometimes conflicting judgments people hold about a robot's capabilities and moral responsibility. This dual approach—validating existing theoretical scales while applying a flexible probabilistic framework—seeks to strengthen both the conceptual and empirical foundation for understanding robotic moral agency.

Applicability

Establishing *when*, *how*, and *why* robots should be considered moral agents is key to designing ethical, trustworthy, and socially integrated autonomous systems. As robotic systems become increasingly embedded in healthcare, transportation, domestic life, and warfare, clear and philosophically grounded frameworks for assessing moral agency are essential for:

•**Ethical system design** – Informing development standards that consider moral responsibility in autonomous decision-making.
•**Policy and regulation** – Guiding legal and policy frameworks about liability, accountability, and the role of autonomous systems in society.
•**Public trust and perception** – Understanding how people attribute responsibility to robots shapes acceptance, oversight, and ethical deployment.
•**Human-robot interaction (HRI)** – Enabling nuanced models of interaction where moral expectations are calibrated to robot capabilities.



As robots become more autonomous, understanding agency becomes more important.

Autonomy, Adaptability, Interactivity Scales (AAI) and Moral Agency

The basis by which this project characterizes Moral Agency for non-human agents is based off of Floridi and Sanders 2004 "On the Morality of Artificial Agents". This paper lays a philosophical and psychological framework in which it is proposed that Moral Agency of artificial agents can be measure through three axioms:

- Autonomy
  - *Ability to change state without stimulus*
- Adaptability
  - *Ability to change the "transition rules" by which state is changed*
- Interactivity
  - *Response to stimulus by change of state*

For an artificial agent to be considered "Morally agent" it must meet all three of these metrics.

Methodology

Data Collection

Participants are shown a series of videos in which a robot performs activities and goes through interactions meant to inform a viewer's opinions on the different AAI moral agency scales. After each video, viewers are asked to respond to a set of survey questions where they will rate the robot's performance on each subscale. After this battery of surveys, respondents are then proposed a scenario where the robot has caused some form of harm to humans and are asked to rate the Moral Agency/Responsibility that should be attributed to the robot in the proposed scenario.

The Nao Robot is the platform being used in the videos as it is excellent at displaying more subtle behaviors.

Analysis Method

The survey responses are taken using the Likert scale, so the data is adjusted using a Basic Probability Assignment (BPA) function to assign probabilities between 0-1 for three different hypothesis subsets:

{[Agree],[Disagree],[Uncertain]}

Then, using a logical combination rule, probability across AAI scales is combined leaving a 3 part metric for every participant's survey, which can be compared to their responses to their rating of Moral Agency.

Survey response (Likert) → BPA → Combined AAI belief → Compared to agency rating

What is Dempster Shafer theory? (DST)

Dempster-Shafer Theory (DST) is a mathematical framework for reasoning under uncertainty. Unlike traditional probability theory, DST allows us to express both belief and ignorance when making decisions based on incomplete or ambiguous data.

Key Concepts

Frame of Discernment (Θ):
The set of all possible outcomes or hypotheses (e.g., $\Theta = \{r_1, \text{not\_}r_1\}$).

Basic Probability Assignment (m):
A function that assigns a value between 0 and 1 to subsets of Θ, representing the probability attributed to that subset.

Belief (Bel):
The total support for a proposition, including all subsets entirely within it.

Plausibility (Pl):
The total support not contradicting a proposition.

Rules of combination

Dempster Shafer theory allows for combination of data through rules of combination. There are a multitude of rules of combination under DST, the most used being Dempster's Rule.

Given two belief mass functions, m_1 and m_2, the combined mass $m(A)$ for a proposition A is calculated as:

$$m(A) = \frac{1}{1-K} \sum_{B \cap C = A} m_1(B) \cdot m_2(C)$$

*Where K is the **conflict coefficient**, defined as:*

$$K = \sum_{B \cap C = \emptyset} m_1(B) \cdot m_2(C)$$

Logical Combination & Uncertain Implication

In addition to numerical rules of combination, researchers Tang and Núñez introduced rules for combination inside DST following logical frameworks and operations (AND, OR, NOT, etc.). In this project, Logical combination, specifically the "AND" rule is important for combining probabilities across the three sections autonomy, adaptability, and interactivity.

Tang's Rule:

$$m(x) = \sum_{A \cup B = x} m_1(A) \cdot m_2(B)$$

Núñez's Rule:

$$m(x) = \sum_{A \cap B = x} m_1(A) \cdot m_2(B)$$

$$m(\bar{x}) = \sum_{A \cap B = \bar{x}} m_1(A) \cdot m_2(B)$$

$$m(\Theta_{\varphi_1,x} \times \Theta_{\varphi_2,x}) = \sum_{A \cap B = \Theta} m_1(A) \cdot m_2(B)$$

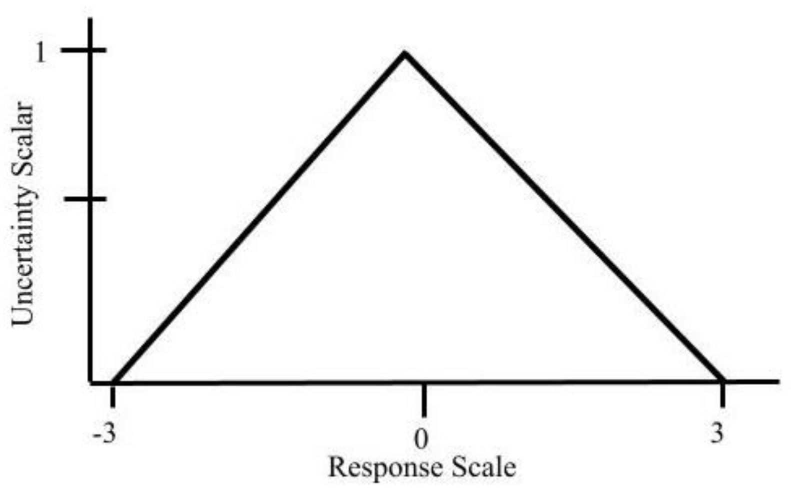
Analysis

Existing measures of moral agency often conflate ontological concepts (e.g., being an agent) with psychological constructs (e.g., appearing agent-like), leading to confusion about what kind of "agency" robots possess. To address this, this project adopts **Floridi and Sanders' AAI framework**, which better aligns with formal philosophical theories, enabling a more rigorous evaluation of robotic systems.

Understanding Moral agency is a blurry area, and **DST** offers a flexible mathematical framework that captures degrees of belief, ignorance, and conflicting evidence, making it ideal for studying moral agency in complex systems.

- **Models uncertainty directly**
- **Supports logical aggregation**
- **Generates richer statistical outputs**

Basic Probability Assignment



This shows an example of how probability is initially assigned to the uncertainty set doing my basic probability assignment. As the responses lean more toward the extremes, probability assignment becomes less ambiguous.

Interactive	Autonomous	Adaptable	Examples
No	No	No	Rock
No	No	Yes	?
No	Yes	No	Pendulum
No	Yes	Yes	Closed ecosystem, solar system
Yes	No	No	Postbox, mill
Yes	No	Yes	Thermostat
Yes	Yes	No	Juggernaut[5]
Yes	Yes	Yes	Human

An Excerpt from Floridi and Sanders paper, showing examples of different systems and their respective traits.

Acknowledgements

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, *14*(3), 349–379. https://doi.org/10.1023/b:mind.0000035461.63578.9d

New Scientist. (2018). *Workplace robots have had little impact on jobs in Germany* [Photography]. https://www.newscientist.com/article/2148924-workplace-robots-have-had-little-impact-on-jobs-in-germany/

Nunez, R. C., Ranga Dabarera, Matthias Scheutz, Briggs, G., Bueno, O., Kamal Premaratne, & Murthi, M. N. (2013). DS-based uncertain implication rules for inference and fusion applications. *International Conference on Information Fusion*, 1934–1941.

RNZ. (2018). *Robots - Nao and the Bristlebots* [Photography]. https://www.rnz.co.nz/national/programmes/ourchangingworld/audio/2018653987/robots-nao-and-the-bristlebots

Tang, Y., Hang, C.-W., Parsons, S., & Singh, M. P. (2012). Towards Argumentation with Symbolic Dempster-Shafer Evidence. *Frontiers in Artificial Intelligence and Applications*. https://doi.org/10.3233/978-1-61499-111-3-462