Human Factors in Robots, Drones and Unmanned Systems, Vol. 93, 2023, 161–170
https://doi.org/10.54941/ahfe1003759

Leveraging Manifold Learning and Relationship Equity Management for Symbiotic Explainable Artificial Intelligence

Sourya Dey, Adam Karvonen, Ethan Lew, Donya Quick, Panchapakesan Shyamshankar, Ted Hille, Matt LeBeau, and Eric Davis

Galois, Inc, Portland, Oregon, 97204, United States

ABSTRACT

Improvements in neural methods have led to the unprecedented adoption of Artificial Intelligence (AI) in domains previously limited to human experts. As these technologies mature, especially in the area of neuro-symbolic intelligence, interest has increased in artificial cognitive capabilities that would allow a system to function less like an application and more like an interdependent teammate. Next-generation AI systems need to support symbiotic, human-centered processes, including objective alignment, trust calibration, common ground, and the ability to build complex workflows that manage risks due to resources such as time, environmental constraints, and diverse computational settings from super computers to edge devices and autonomous systems. In this paper we review current challenges in achieving Symbiotic Intelligence and introduce solutions in the form of Artificial Executive Functions aimed at solving these challenges. We present our work in the context of current literature on self-aware computing and present basic building blocks of a novel, open-source, AI architecture for Symbiotic Intelligence. Our methods have been demonstrated effectively in both simulated crisis and design problems and during the pandemic. We argue our system meets the basic criteria outlined by DARPA and AFRL providing: (1) introspection via graph-based reasoning to establish expectations for both autonomous and team performance, to communicate expectations for interdependent co-performance, capability, and understanding of shared goals; (2) adaptivity through the use of automatic workflow generation using semantic labels to understand requirements, constraints, and expectations; (3) self-healing capabilities using after-action review (AAR) and co-training capabilities; (4) goal oriented reasoning via an awareness of machine, human, and team responsibilities and goals; (5) approximate, risk-aware, planning using a flexible workflow infrastructure with interchangeable units of computation capable of supporting both high fidelity, costly, reasoning suitable for traditional data centers, as well as in-the-field reasoning with highly performable surrogate models suitable for more constrained edge computing environments. Our framework provides unique symbiotic reasoning to support crisis response, allowing fast, flexible, analysis pipelines that can be responsive to changing resource and risk conditions in the field. We discuss the theory behind our methods, practical concerns, and our experimental results that provide evidence of their efficacy, especially in crisis decision making.

Keywords: Symbiosis, Human-machine teaming, Human-robot interaction, Interdependence, Artificial cognition

AI and machine learning enabled systems have become a ubiquitous part of modern life. Spurred initially by the DARPA PAL (Personalized Assistant that Learns) and CALO (Cognitive Assistant that Learns and Organizes) programs (Gouin, 2012) these Intelligent Software Assistants (ISAs) are now standard on most smart phones, home assistants, intelligent thermostats, and other home automation technologies. While these systems have greatly enhanced the productivity of their users, they fall short of their promises, delivering capabilities that are more focused on single interaction workflows with their human co-performer, lacking truly interactive capabilities.

We discuss a novel framework for constructing machine co-performers (MCPs) with artificial cognitive capabilities designed to recreate key portions of executive function believed vital for generating symbiotic AI, extending current capabilities beyond mere appliances or automated assistants, and enabling truly interdependent workflows, generating MCPs that have **mission-based** models to account for **sense-of-self** and **theory-of-mind** in interdependent workflows.

- **Sense-of-self** – Self-aware computing was the subject of a DARPA workshop in 2004 (Amir, 2007). In humans self-awareness is a poorly understood process relating to knowledge of one's permanent aspects, relationships to others, sensory experiences, beliefs, desires, intentions, and goals. In computing the topic of self-awareness has generally been discussed as **self-monitoring**, where a computer system monitors, evaluates, and intervenes on its internal processes in a purposeful way; and **self-explanation** or **metacognition** where a system can accurately recount and justify the actions and decisions it has made. In this paper we focus on a **mission-bound sense-of-self** which enables self-monitoring and explanatory behaviours associated with an explicit mission or purpose for a **human-machine team**. In human-machine teaming we argue that sense-of-self is important for interdependence, as it allows a machine co-performer (MCP) to act with similar self-understanding of its actions and goals as a human teammate.
- **Theory-of-mind** – In addition to sense-of-self, an equally important symbiotic property is theory-of-mind. In psychology, the notion of theory-of-mind implies the existence of an explicit model of other co-performers in an environment paired with the knowledge that those co-performers have mental states; that these states may be different from the state of the agent with theory-of-mind; allowing judgement and inference on the **mission-bounded** behaviours of human co-performers. While not outlined explicitly by DARPA in their discussion of self-aware AI, theory-of-mind, and other notions of machine awareness of the contexts, emotions, and goals of their co-performers is vital to achieving symbiotic workflows (Saracco, 2021).

In their 2009 report (Agarwal, 2009) a set of desired capabilities for so-called "Self-Aware AI" are presented consisting of five stated goals and functionalities:

1. *It is INTROSPECTIVE or SELF-AWARE in that it can observe itself and optimize its behavior to meet its goals.*
2. *It is ADAPTIVE in that it observes the application behavior and adapts itself to optimize appropriate application metrics such as performance, power, or fault tolerance.*
3. *It is SELF HEALING in that it constantly monitors its resources for faults and takes corrective action as needed. Self-healing can be viewed as an extremely important instance of self-awareness and adaptivity.*
4. *It is GOAL ORIENTED in that it attempts to meet a user's or application's goals while optimizing constraints of interest.*
5. *It is APPROXIMATE in that it uses the least amount of precision to accomplish a given task. A self-aware computer can choose automatically between a range of representations to optimize execution – from analog, to single bits to 64-bit words, to floating point, to multi-level logic.*

We present a roadmap to achieving these core goals through the development of **Artificial Executive Function**, along with our core framework for enabling these functions in AI and machine learning systems. The so-called **executive functions** of the human brain, sometimes referred to as cognitive control, are a set of cognitive processes evident in the human brain that allow individuals to understand and modulate their behaviors through monitoring, the exercise of control, inhibition, and are built on underlying substrates such as working memory. Current belief is that higher-order executive functions are a requirement for planning, fluid intelligence, abstract reasoning, and problem solving (Diamond, 2014). While there is some disagreement in the literature on the nature and exact taxonomy of executive functions, recent research has somewhat clarified the confusion by demonstrating the experimental separability of **mental set shifting, information update and monitoring,** and **inhibitory control** (Miyake, 2000). In their paper Miyake et al. were able to show that individuals with impaired executive function development had measurable deficiencies in these areas that directly impacted problem solving abilities.

Our framework enables artificial executive functions, achieving many of DARPA's goals for self-aware AI through its use of **working memory** that stores a history of co-performance outcomes with individual human co-performers, tracking expectations for co-performance generated using its training data and prior records of engagements. The framework allows an MCP to observe prior independent behavior, recorded joint behaviors and team performance outcomes, and simulated autonomous action in similar contexts to understand skill and objective gaps in the team's co-performance that were either corrected by the MCP intervening on human behaviors or actions or corrected by the human intervening on the machine's behaviors and actions.

TRUST AND CO-TRAINING IN HUMAN-MACHINE TEAMS

In order to address increasing threats from adversaries, warfighters must be partnered with intelligent systems capable of automating and augmenting

human capabilities. Traditional models of these human-machine teams have autonomous partners, and human partners train independently over the skills needed to accomplish a task or mission. We argue that in order for these teams to be truly interdependent it is necessary for them to co-train as well as co-perform. The Air Force has long documented the advantages of co-training for teams during co-performance situations (Stone, 1999). When understanding how team composition impacts success or failure on goals during co-performance, prior co-training was found to be among the most influential for positive outcomes. Teams that train together tend to have better calibrated trust in the skills and limitations presented by co-performers; additionally, co-training seems to help humans develop an intuition as to the cognitive flexibility of their partners, and the ability of their partners to acquire new skills in the field. Co-training between humans and machines, we theorize, not only helps to develop these capabilities in human co-performers, but also provides Symbiotic AI with a chance to build models of their co-performer's intent, capabilities, and most importantly, their emotional state.

Co-training in repeated exercises between human and machine partners, during complex training scenarios that provide adequate models of anticipated missions, allows for the development of bi-directional trust between the human and machine teammates. Once established, we use these models to measure the **emotional equity** and outcomes from interdependent co-performance and to establish, maintain, and repair trust with a human co-performer. We argue that these interdependence relationships are the key way that trust is established, maintained, and repaired amongst human performers, and by providing a model of trust in this context, we can enable more effective machine teammates, and exercises which establish appropriate levels of trust in the human performer.

TRUST MODELING

When developing a relationship with a machine teammate, or other piece of necessary technology, most human performers enter the relationship with some level of inherent bias with respect to trust. This inherent bias is usually **miscalibrated**, resulting in **under trust** of the system, or **over trust** of the system. While this initial bias is developed independently, and thus cannot be controlled by a Symbiotic co-performer, this initial trust assessment can be modified by the MCP as they co-train with their human counterparts. It is a central thesis of ours that trust is not static (Feltovich, 2004), but can be directly influenced through the experience gained by teaming with machine counterparts.

MCPs built with our framework use a modified version of the Trust model from (Akash, 2017) which was itself an extension from (Jonker et al. 1999, Jonker et al. 2004). The MCP builds an expectation for the current trust calibration based on its own expectations, and outcomes, and extends the signal detection classification approach presented in (De Visser 2020), maintaining a model of the current relationship equity, the current level of trust displayed by the human (as estimated by behavioral, self-report, and economic bet

measures from telemetry), and determines if its expectations of performance represent a **hit, false alarm, miss,** or **correct rejection**, as in De Visser. We extend the De Visser model by including expectations for co-training, and engaging in similar emotionally vulnerable communication patterns during After Action Review (AAR) not only about performance (to calibrate trust in performance), but also about co-training improvements and future outcomes, communicating its goals in co-training, the outcomes, and how these outcomes shift expectations for the MCP if capabilities improved, degraded, or remained roughly equivalent, shown in Figure 1.

While in De Visser these classifications are used to characterize behaviour, we utilize our extended classification to set the **mood state** of the MCP, indicating the understanding of the current trust process and the self-healing adaptive action that is to be taken for correction. These mood states are used to seed our After Action Review process, selecting the subjects to discuss, and the lexical interventions to take with the human co-performer to modulate trust into a more calibrated state.

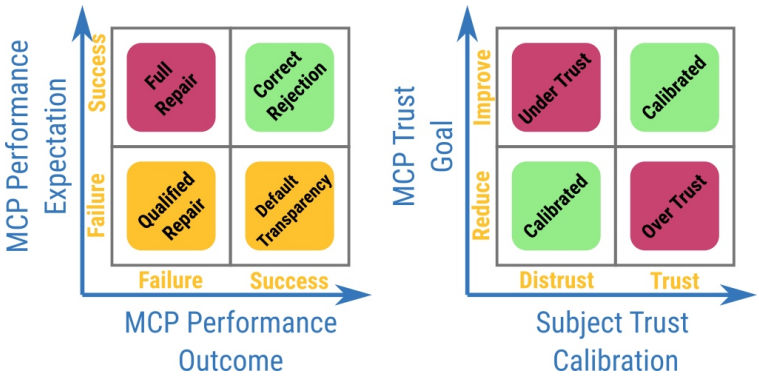

Figure 1: The MCP plans actions for trust calibration and co-training on the basis of metacognitive analysis of prior performance, and its expectations for that performance based on prior engagements and training. Based on those outcomes, it then establishes a trust goal, and compares subject trust calibration signals to its goal to plan its AAR engagement.

MODELS OF JOINT ACTIVITY

Joint activity between human and machine co-performers are modelled in our framework using a formalism of human-readable box-diagrams that describe a workflow beginning with the data sources available to the MCP, and ending in the goal states the MCP is seeking to achieve as a sort of "blueprint" or template for the joint activity. The MCP then may select functions from its inventory of computational capabilities in order to seek to complete this abstract workflow.

We provide automatic workflow generation through our open-source ORCA framework (Cowger, 2022), producing models of joint activity that are both actionable, and explicitly represent points of human-machine interdependence, interaction, and models for activity and task coordination. The

ORCA system is flexible in orchestrating tasks, and based on a container execution system (currently supporting Docker, Kubernetes, and other compatible standards). This allows not only the orchestration of existing capabilities, but the containerization and utilization of novel capabilities, including those synthesized by the Symbiotic AI itself using surrogate modeling, or through further training of its neural components.

The use of workflow execution allows our MCP to observe and optimize its own behaviors through co-training and after-action review; adapting its behavior to changes in the difficulty, mission and task complexity and difficulty, as well as trust signals and capabilities displayed by its human co-performers. The ORCA framework also enables self-healing capabilities of our AI systems, applying co-training and trust calibration procedures automatically in response to failures or gaps in performance. The system is goal-oriented in nature, ORCA workflows are fully compatible with joint-activity graph representations of interdependent activity (Johnson, 2021) allowing the construction of workflows automatically from expectations and inputs provided by the user. Individual tasks in the workflow can be replaced with approximate reasoning through the Type 1 and Type 2 reasoning (Evans, 2013) provided by our surrogate model generation functions. Computationally expensive, solutions can be approximated with our own prior work on Koopman theory methods using deep learning (Dey, 2022).

Figure 2a shows an example of the initial state of joint activity graph construction in which the MCP receives information on its inputs and outputs in the form of semantically labeled data cubes with annotations on constraints, requirements, and expectations on the error and confidence of the outcomes; an intermediate, partial, workflow graph is shown in Figure 2b. In this stage, we show a partially constructed workflow graph in which the MCP has decided to attach a filter function to split its workload into three parallel paths, and begin identification tasks, with the outcomes of these processes indicated in the new graph, and then finally the completed joint activity graph. The MCP constructs this graph in such a way to maximize its confidence in achieving its tasks, while minimizing risk of failure from its inventory of existing computational functions. Each function is modelled as a container with both semantic and type labels on its input set, output set, and on the function itself

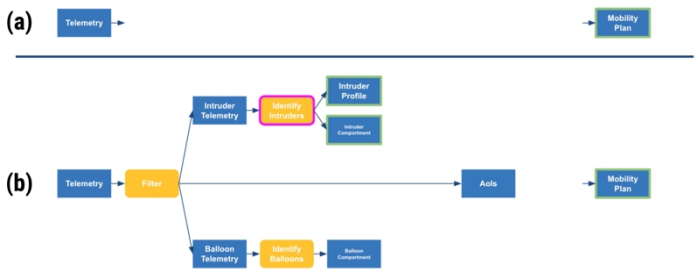

Figure 2: Example of a partially assembled joint activity graph in ORCA. In (a) we see the initial graph, defined by the initial input, and goal for the output. In (b) the MCP has created a partial workflow, working to satisfy the subject's request.

which is used by the MCP to determine appropriate actions. Each container's semantic labels serve as expectations on the output from the function, and its compatibility with other units of automated reasoning.

Figure 3 shows the final constructed joint activity graph with labels for tasks that can be shared with the human co-performer and for which some level of human feedback is anticipated. In this example the tasks deal directly with the sensing of a contested airspace, and air mobility planning to avoid loss of separation with intruders and aerostatic obstacles, while simultaneously attempting to achieve proximity to areas of interest in the environment. Tasks 1–4 deal with Identification of Intruders, Intruder Loss of Separation estimates, Balloon Loss of Separation estimates, and final mobility planning, and can all be delegated to the MCP by a human, or accomplished by a human directly. Both human and MCP tasks end in data cubes with the same type signatures, making them directly compatible, though our prior work in automatic work flow generation also means our ORCA framework is able to adapt units, assumptions, and representational differences across different users, if needed, in a real engagement.

These joint activity graphs can also readily be converted into fault-trees, as shown in Figure 4, using an automatic algorithm. The MCP uses this capability to conduct after action review, and to apply its metacognitive functions to

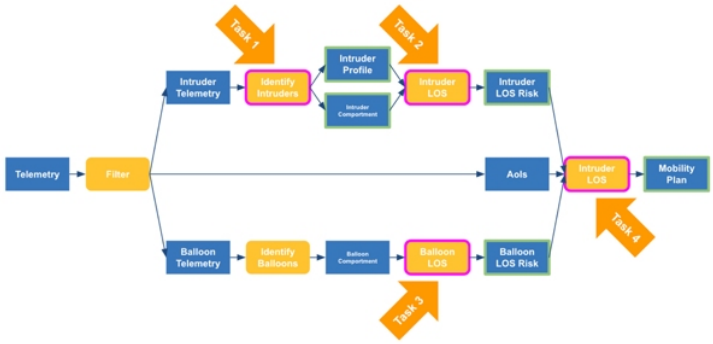

Figure 3: A final joint activity graph constructed for the MCP and subject's interdependent workflow. Tasks highlighted with arrows can be executed by either the MCP, or human co-performer.

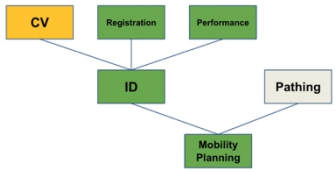

Figure 4: ORCA workflows are not only compatible with joint activity graphs, they can also be directly converted into fault trees for analysis, blame assignment, and after-action review.

determine co-training potential with its working memory. Before subsequent performance, the MCP establishes expectations on:

- Individual performance - The MCP attempts to judge its own skill level by looking at training and validation data, and then estimating its performance outcomes, likely score, and ability to score while avoiding loss of separation without a human co-performer. This sets a baseline for expectations on performance.
- Joint performance - The MCP examines its working memory (part of its nascent Artificial Executive Functions) for prior engagements (if any) with the current user and establishes an expectation of joint performance, as either better than, worse than, or roughly equivalent to its own individual performance.
- Co-training outcomes for each of its functions - The MCP has an understanding of diminishing returns in its own learning functions, and establishes expectations on the result of co-training of any of its functions.

AFTER ACTION REVIEW AND MANIFOLD LEARNING

Simply being aware of its actions is not enough to meet DARPA's stated requirements for "self-aware AI", nor is it a full exercise of the artificial executive functions we have been developing. Equally important is the ability to **self-explain** the actions, behaviors, and reasoning displayed by an MCP. Our framework implements After Action Review (AAR) capabilities as a communication tool between humans and machine co-performers. During AAR, each team member is able to review the team's joint performance and discuss areas for improvement. This act of reflection facilitates trust transfer and calibration, which are crucial to both the performance of the team and the measurement of trust calibration. Foundational methods of communication gleaned from the domain of eXplainable AI (XAI) have been employed to aid in the interpretability and explainability of the MCP's reasoning and decision making (Klein et al, 2021). Our methodology aims at minimizing the explainability gap by utilizing natural language-like structures to explain the AI's performance based on past training experience in conjunction with the relevant task. Unlike methods utilizing large-language models (LLMs) (Bang, 2023), we take an approach of using templated dialogue that is populated by explanations from embedded manifold representations of decisions and outcomes using Self Organizing Maps. Manifold learning enforces Euclidean properties, while also performing dimensionality reduction, on complex decision spaces.

As shown in Figure 5, our framework allows MCPs to embed its decisions in explainable manifolds, representing reasoning as a series of feature planes important to an embedding. Explainable features from symbolic reasoning systems as well as latent features from neural reasoning systems can be composed in these manifolds, which are then selected to minimize quantization, topographic, and classification error. The resulting embedding allows the MCP to select the most relevant human decision features for a classification, and to help human co-performers understand trade-offs in design and decision spaces during after action review. Furthermore, remediative actions

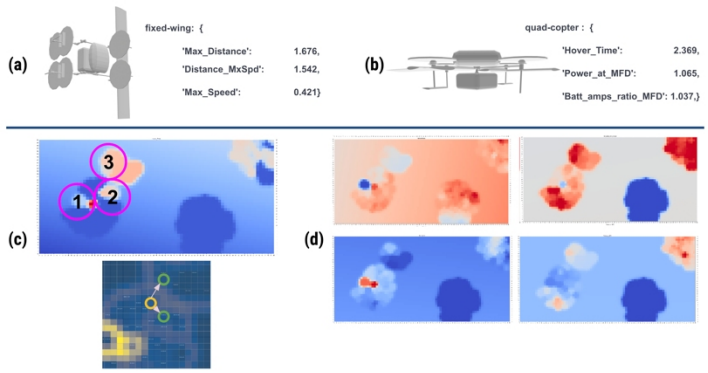

Figure 5: Manifold learning allows not only dimensionality reduction, but also explainable feature extraction for perceptual alignment as shown in (a) and (b). Individual feature maps can then be used to understand decision trade-offs (c, d) when compared to commander's intent.

from the human co-performer can likewise be embedded in the manifold, allowing an MCP to store and adapt to semantically labeled criticisms from its human performer.

CONCLUSION

In this paper we have presented a summary of novel functions for Symbiotic Artificially Intelligent systems enabled by our framework, and their utility in human-machine co-performance. During the coming months we will be testing these functions and their coverage of artificial executive function in human-subjects research to establish their efficacy for establishing, maintaining, and repairing trust; and improving the capabilities of human-machine teams to work effectively on interdependent tasks.

REFERENCES

Agarwal, A., Miller, J., Eastep, J., Wentziaff, D. and Kasture, H., 2009. *Self-aware computing*. MASSACHUSETTS INST OF TECH CAMBRIDGE.

Akash, K., Hu, W. L., Reid, T. and Jain, N., 2017, May. Dynamic modeling of trust in human-machine interactions. In *2017 American Control Conference (ACC)* (pp.1542–1548). IEEE.

Amir, E., Anderson, M. L. and Chaudhri, V. K., 2007. *Report on DARPA workshop on self aware computer systems*. SRI International Menlo Park United States.

Sam Cowger, Sourya Dey, Ethan Lew, Panchapakesan Shyamshankar, Ted Hille, Eric Davis, 2022. *Orca: Orchestrating Symbiotic Intelligence for Agile and Adaptable, Crisis Response Decision Making*. The Chesapeake Large Scale Accelerator Conference, Annapolis, MD.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W. and Do, Q. V., 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *arXiv preprint arXiv:2302.04023*.

Diamond, A., 2013. Executive functions. *Annual review of psychology*, *64*, pp. 135–168.

Dey, S. and Davis, E., 2022. DLKoopman: A deep learning software package for Koopman theory. *arXiv preprint arXiv:2211.08992*.

De Visser, E. J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S. and Parasuraman, R., 2012, September. The world is not enough: Trust in cognitive agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 56, No. 1, pp. 263–267). Sage CA: Los Angeles, CA.

Evans, J. S. B. and Stanovich, K. E., 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, *8*(3), pp. 223–241.

Feltovich, P. J., Bradshaw, J. M., Jeffers, R., Suri, N. and Uszok, A., 2004. Social order and adaptability in animal and human cultures as analogues for agent communities: Toward a policy-based approach. In *Engineering Societies in the Agents World IV: 4th International Workshops, ESAW 2003, London, UK, October 29-31, 2003. Revised Selected and Invited Papers 4* (pp. 21–48). Springer Berlin Heidelberg.

Gouin, D., Lavigne, V. and Bergeron-Guyard, A., 2012. Human-computer interaction with an intelligence virtual analyst. *Proceedings of Knowledge Systems for Coalition Operations, IHMC, Pensacola, FL*, pp. 1–5.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J. and Parasuraman, R., 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, *53*(5), pp. 517–527.

Hoff, K. A. and Bashir, M., 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), pp. 407–434.

Hoffman, R. R. and Deal, S. V., 2008. Influencing versus informing design, part 1: A gap analysis. *IEEE Intelligent Systems*, *23*(5), pp. 78–81.

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., Van Riemsdijk, M. B. and Sierhuis, M., 2014. Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, *3*(1), pp. 43–69.

Johnson, M. and Vera, A., 2019. No AI is an island: the case for teaming intelligence. *AI magazine*, *40*(1), pp. 16–28.

Jonker, C. M. and Treur, J., 1999. Formal analysis of models for the dynamics of trust based on experiences. In *Multi-Agent System Engineering: 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'99 Valencia, Spain, June 30–July 2, 1999 Proceedings 9* (pp. 221–231). Springer Berlin Heidelberg.

Jonker, C. M., Schalken, J. J., Theeuwes, J. and Treur, J., 2004. Human experiments in trust dynamics. In *Trust Management: Second International Conference, iTrust 2004, Oxford, UK, March 29-April 1, 2004. Proceedings 2* (pp. 206–220). Springer Berlin Heidelberg.

Klein, G., Hoffman, R. and Mueller, S., 2021. Scorecard for self-explaining capabilities of AI systems.

Mayer, R. C., Davis, J. H. and Schoorman, F. D., 1995. An integrative model of organizational trust. *Academy of management review*, *20*(3), pp. 709–734.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A. and Wager, T. D., 2000. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive psychology*, *41*(1), pp. 49–100.

Saracco, R., Grise, K. and Martinez, T., 2021. The winding path towards symbiotic autonomous systems. *Philosophical Transactions of the Royal Society A*, *379*(2207), p. 20200361.

Stone, B. M., Turner, K. L., Rue, R. C. and Mitchell, J. L., 1999. *A Task-Based Approach to Analyzing Processes*. METRICA INC BRYAN TX.