Artificial Intelligence and Social Computing, Vol. 163, 2025, 158–167
https://doi.org/10.54941/ahfe1006047

Humans and AI Based Communication and Reasoning in Complex Adversarial Domains

James Llinas

State University of New York at Buffalo, Buffalo, N.Y., 14260, USA

ABSTRACT

It is generally agreed that trust is best conceptualized as a multidimensional psychological attitude involving beliefs and expectations about the trustee's trustworthiness derived from experience and interactions with the trustee in situations involving uncertainty and risk. It has to do with the notion of a willing exposure to risk and an agent willing to be vulnerable to "the other". In this paper we explore ideas about credition, the interdisciplinary process of believing, and how communicating agents get to believe each other, how issues of uncertainty enter into the issue of believability, and how belief and consciousness also interplay. The paper also addresses epistemological issues related to reasoning and analytical approaches that integrate multidimensional perspectives (labeled "epistemic pluralism") for complex adversarial domains such as those involved with modern and future intelligence analysis.

Keywords: Communication, Reasoning, AI, Belief, Epistemology

CREDITION

Credition, the processes of believing, is a fundamental brain function that enables a human being to trust his/her inner probabilistic representation(s) of some current world condition, as perhaps communicated to the human by an AI system or developed from the gathering of evidence. In this situation, one could ask: "How did the AI system develop its own belief of the world condition, and how was that belief justified?" That is, why should the human *believe* the assertions of an AI system and similarly why should the AI system *believe* the message communicated by a human? How do these belief operations work exactly, and how is it that each agent comes to a point where they are willing to act on the basis of the communicated proposition? If we hark back to theories of knowledge, knowledge most often represented as "justified true belief" (in something we are asserting we "know"), then assertions of having knowledge of some world condition means we believe that state exists, and we have somehow justified our assertion. Sensibly all models of credition assert that it is based on neural processes, including perception and valuation of objects and events in the physical and social environment. In the context of a dynamic real-world being assessed by an AI system and a human in a shared environment, managing the balance of

exploration (i.e., the examination of *alternative* beliefs) versus *exploitation* (the use of an *existing* beliefs to make a decision) can be critical, especially in any situation involving a sense of urgency or in adversarial settings where balancing such alternatives may be critically important. Timing is also an issue; delays due to contemplation about belief can lead to serious consequences or can also be a result of believing and acting too soon. This balance can be looked at as the distinction between the concepts of estimation and judging. Estimation is the type of calculation provided by current AI systems, that yield an assessment; judging is the human kind of decision making, based on the evaluation of embodied knowledge and beliefs, and the human's *conscious attitude*. This balance, which is also key to the performance of Machine Learning (ML) or Reinforcement Learning (RL) algorithms (Lumbreras, 2022), can be seen as an essential feature of *managed belief formation and update* among/between AI systems and humans that, importantly, depends on various characteristics of the involved humans and AI processes. Thus, there is a challenge in any bi-lateral AI-Human system environment involving achievement of some common purpose or goal to achieve a *common state of belief* at any moment.

Belief is a complex topic. According to (Vestrucci et al., 2021), there are at least six models of how beliefs function. These models have a variety of frameworks to include self-organizing systems and others, including complex system forms. The credition function depends on, and acts along, two different dimensions: cognition and emotion. Inclusion of an emotional factor to a belief model raises an interesting question currently being studied in some AI circles of AI systems having a *conscious attitude*; we make some remarks about consciousness and AI later. The synergies between understanding belief formation and artificial intelligence suggests that "AI still has plenty of unexploited metaphors that can illuminate belief formation" (Lumbreras, 2022). In addition, acknowledging that AI should be integrated with our belief processes (e.g., the capacity to reflect, rationalize) makes it possible to focus on more promising lines such as Interpretable Machine Learning (Lumbreras, 2022). Indeed, if we hark back to two recent surveys on Interpretable AI, we find that those assessments conclude that "there is a lack of agreement on what constitutes a comprehensible or understandable explanation" (Alangari, 2023), and "The literature currently being generated on interpretable and explainable AI can be downright confusing" (Rudin, 2022). Neither of these survey papers specifically includes notions of belief, but it seems clear that even indirectly the goals of these efforts toward Interpretable AI are to convey notions of believability. Clearly, the absence of transparency and interpretability for the operations of, and the decisions for, these systems indicate a deficiency that can have severe consequences in all domains but especially in those domains such as medical diagnosis, financial decision-making, and adversarial military operations where valuable resources and lives are at stake. These efforts to provide interpretable AI systems are not only needed for today's AI problems addressing the main error modes of overfitting and biases (not easy to detect in today's black box AI processes) but for the more ambitious applications of AI, and to move toward achieving AI systems that "can be understood as

a machine that supports the formation and valuation of beliefs in the human and can be understood metaphorically as a belief-machine itself" (Lumbreras, 2022). For these reasons, it is extremely interesting to examine AI as to how it may enable the credition process because understanding how AI works can give users insights to inform our hypotheses about how credition works. In parallel, acknowledging that AI should support belief formation helps to design it better and make this support as effective as possible.

BELIEF AND UNCERTAINTY

At the heart of virtually all processes attempting to understand the world is uncertainty. Uncertainty reasoning and quantification, largely as related to both estimation of situational states and in decision-making have been studied for many years in various AI domains (e.g., belief/evidence theory, game theory, data fusion, and machine/deep learning). In the AI community, a variety of belief or evidence theories have a long history of application in studying reasoning and decision-making under uncertainty. The development of Machine Learning and Deep Learning (ML/DL) algorithms have mostly considered two common uncertainty types, *aleatoric uncertainty* and *epistemic uncertainty*, for decision-making. In their survey of uncertainty and Deep Learning (DL), (Guo et al., 2023) argue that it is critical to quantify diverse types of uncertainty caused by different root causes, which may lead to the formation of different hypotheses and different levels of belief for a decision-maker. For example, recent studies have combined different belief models with DL process models to quantify different uncertainty types about the predictions made by a DL model. Aleatoric (aka statistical) uncertainty refers to the notion of randomness, that is, the inherent variability in the outcome of an experiment which is due to inherently random effects. Epistemic (aka systematic) uncertainty refers to uncertainty caused by a lack of knowledge, i.e., to the epistemic state of the agent. As opposed to aleatoric uncertainty, epistemic uncertainty can in principle be reduced on the basis of additional information. For ML-based AI, epistemic uncertainty usually results from the uncertainty in the knowledge that justifies the model weights. A review of how these different types of uncertainty influence ML-based AI processes is shown in (Senge et al., 2014). There are a number of statistical modeling frameworks that connect notions of belief to assertions of numerical degrees of belief in a proposition; clearly this paper cannot review all of those models and their notions of belief.

BELIEF AND CONSCIOUSNESS

Belief formation in human beings involves elements of intuition, empathy, and creativity, asserted to be related to human consciousness. However, for AI to progress to more complicated tasks requiring intuition and empathy, it must develop capabilities such as metathinking, creativity, and empathy similar to human self-awareness or consciousness. There are many papers offering ideas as to how AI capability might evolve to such a level. In (Esmaeilzadeh and Vaezi, 2021), the requirements for the emergence of consciousness in AI are explored; two camps seem to exist, one asserting the

need for biological awareness and the notion of consciousness experienced by humans, the other saying consciousness can be achieved through neurological processes and thus through advanced computational capabilities. We select this paper and its views because of the theme of this paper on human-AI communication. This is because (Esmaeilzadeh and Vaezi, 2021) argue that consciousness is a social phenomenon, wherein agents become aware of their consciousness *as a result of* communication with another agent.

Human-AI Communication

First of all, consistent with the themes of this paper, considering an AI capability as a "communicator" means that AI technology is stepping into a role previously restricted to humans, and this imputes (much like what was said above) the social dimension of an AI agent. This means adding an adaptive capability regarding context, the properties of the "other", and the specific contents of the messages. That is, such capability means an ability to learn and adapt to a human or other AI partner and adjusting communication interactions accordingly. It also raises the question of what is meant by "meaning" and "making meaning". For the current context, we take the definition from (Solum, 2014) that is based on a communicative, inter-agent setting. In that work, meaning refers to *communicative content*— the content conveyed by the text, given the context in which it was authored. But important here is for each recipient to *understand the intent* of the messages, *the "illocutionary" aspect of the messages*, meaning the speaker's *intention* as distinct from what is *actually said* –"*reading between the lines*".

AI and Language

In a book that has generated widespread commentary, Landgrebe and Smith (2022) have a chapter titled "Why machines will not master human language", in which they present a wide range of arguments supporting the implied assertions of their title. There is no space in this paper to review these thoroughly, but we offer a few excerpted remarks relevant to the topics of this session. In the sections having to do with human-machine conversation, they address the reciprocal tasks of

1. the production of the initial utterance of a dialogue, and
2. the maintenance of the dialogue in succeeding utterances.

This dynamic first requires the act of *choosing to produce* an initial utterance, needing the ability to understand the *context* in which the AI-human partners find themselves, and then the maintaining of the dialog that requires considering the role-switching of each agent over time. The first requires an AI agent to understand some situation and to create an appropriate *particular phrasing* for its first remarks. The understanding of this initial utterance first requires the recipient to carry out syntactical analysis and construction, and a complex second step in which context-dependent meaning is derived and assigned to the uttered sentence. Landgrebe and Smith assert the impossibility of AI to contextualize such dialogues across rolling, changing contexts that occur with each utterance, i.e., across the dynamics of a conversation This is because (they argue) the

world knowledge enabling the interpretation of such dialogues, which can be combined in arbitrary form to create many different sorts of contexts, cannot be learned without life experience and (they say) it cannot be mathematically formalized. AI technology cannot decide how to fill in implicit meaning developed as a result of terse language or slang, or of incomplete expressions from a human. Their chapter also remarks that each agent will switch roles as utterers and interpreters as they take turns based on cues if communicating in "human" fashion (i.e., cutting each other short, interrupting, speaking at the same time), and the AI to address such dynamics and ambiguities does not exist. Finally, we cite an important assertion they make regarding mathematical approaches taken by modern AI systems toward modeling language: "the systems which produce human utterances are evolutionary, non-ergodic, driven, and devoid of fixed-boundary conditions (in other words, they are highly context-dependent); from this it follows that to model a conversation by the drawing of sample utterances from a multivariate distribution is impossible: there is no multivariate distribution from which one could draw samples to obtain a stable training set for a stochastic conversation model—and, therefore, there is no adequate retraining either". Further remarks are offered in the chapter and across the book that apply to this paper's theme.

Discussion on Human-AI Communication

This paper has suggested that for almost all applications that will have AI systems and processes interacting with humans, even those that are non-critical and non-urgent, each agent will have to *believe the other* in the same way or in very similar ways that humans come to believe each other. The abilities for, and the methods by, which AI processes and humans communicate beliefs to each other are going to be central issues for realizing AI-human potential synergies in any application. From the review of various research topics examined here that bear on the issues surrounding such capabilities, there is a long way to go before these goals can be achieved. The research reviewed here suggests that AI processes will have to be evolved that have much more human-like capabilities and features, such as consciousness, empathy, and much stronger language capabilities. But if we consider the views of (Landgrebe and Smith, 2022), AI will never have human-like linguistic capabilities, and if so, some assessments need to be made about other ramifications regarding human-like capabilities. But as pointed out in (Esmaeilzadeh and Vaezi, 2021) and remarked above, they suggest that a less-than-natural basis of AI-human communication may be adequate for each agent to develop a sense of consciousness; maybe less-than-natural bases of communication may be adequate to communicate belief as well, and such thoughts may also lead to new models of communication that are departures from the historical, human-based models.

COMPLEX ADVERSARIAL PROBLEMS: THE CHANGING INTELLIGENCE ANALYSIS PROBLEM SPACE

The 2022 Intelligence Community Directive IC 203 (ICD203) sets a pretty high bar for analytical standards for Intelligence Analysis (IA) but it makes no remarks about the underlying epistemic, knowledge-based operations that would result in achieving those standards. In this paper we review and discuss some ideas addressing first the nature of modern and evolving IA problems and their complexities, and then, based on a review of a number of papers, suggest some paths to epistemic solutions for these newest problems.

In today's complex international environment, we are seeing new and still-evolving notions of adversarial interactions to include Cognitive Warfare, Neurowarfare, and Hybrid Warfare, among possibly other still-evolving ideas about the nature and complexities of the modern PS in which IA and policy/decision-making processes will find themselves. There are a number of papers offering a yet wider range of thinking on modern warfare models, e.g., (Berzins, 2020), but again we constrain our discussion because of scope.

It could be argued that these strategies attempt to move warfare into the *complex systems domain* by incorporating multiple interconnected, interdependent warfare lines of operation. Several papers and books address this yet additional issue of how to address the complexity issue in intelligence analysis to move beyond linear, cause-and-effect and inward-looking reductionist analysis. In (Duvenage, 2010), a range of issues are mentioned as regards the need for IA to incorporate aspects of complex systems theories into analytic methods. Among the points made there are: dealing with paradoxical outcomes, discontinuities and tipping points, the irreducibility of systems, and emergent behavior.

COMPLEXITY, NETWORKS, ONTOLOGIES, INFORMATION-HOSTILE ENVIRONMENTS, AND EPISTEMOLOGIES

Complexity

Perhaps the overarching common aspect of these new views of warfare types is the issue that they raise of problem complexity. Complexity concepts are still a collage of principles, methods and concepts, which have not yet been formed into a real coherent framework but most agree they are about *complex relational patterns and non-linear phenomena*, which are not really addressed by Newtonian science. Modern analytically-based methods can be described as employing a "divide and conquer," approach, and they are rooted in the assumption that complex problems are solvable by dividing them into smaller, simpler, and thus more tractable units. Because the processes are "reduced" into more basic units, this approach has been termed "reductionism" and has been the predominant paradigm of science for a long time. Much of the epistemology of IA has followed this type of approach, applying structured analytics and other similar paradigms. Investigating a system in this way implies that the explanatory power derived from understanding its components is sufficient to understand the whole (interacting) system, and that the properties of such components are not

affected by their interaction with their operating environment, i.e., by their operational context. But this is only valid for closed or isolated systems that do not interact with their situated environment. This is particularly true for example with biological systems that are very intimately connected to their host (body) environments. The notion of complexity envelops the very general idea that most phenomena *cannot be tackled in terms of classical reductionism*, that is, they cannot be conceived of as the result of the interaction of their separate parts. The limits of a reductionist approach to studying biological systems are for example discussed in the highly cited paper by (Ahn et al., 2006). Ahn remarks that today's clinical science is fundamentally reductionist, typical procedures involving the search for the failing part; e.g., tumors for cancer, infection for the pathogen. Such methods fail to incorporate contextual factors, and, for example, elderly people often get the same treatment for a problem as young people, as one example of a context-free treatment approach. Ahn points out that "the complex interplay between parts yields a behavior that cannot be predicted by the investigation of the parts". Ahn's paper points out the need for *epistemic pluralism*, which emphasizes the utility of a *multitude of perspectives and types of knowledge, and contextualism*, the ideas of which are developed in (Canali and Lohse, 2024), also referencing the experiences over the Covid-19 period. We elaborate on epistemic pluralism below.

Ontology

A synthesized approach brings us back to ontology. It can be argued that any analytic or epistemological approach needs to have common understanding of the components of the PS; to provide that, most analytic and epistemological methods employ an ontology. In (Smith, 2012), a characterization of ontology is described as "Ontology as a branch of philosophy is the science of "what is", of the kinds and structures of objects, properties, events, processes, and relations in every area of reality". In intelligence, the ontological problem is related to the nature and characteristics of entities that threaten and are threatened. Relatively few ontological structures have addressed the ontology of threats as needed by any IA approach, but our research center developed draft ontologies of threats in the publications of Little and Rogova previously cited. As far back as 2012, (Llinas, 2012) offered a paper on "Framing and Defining New Fusion Strategies and Advanced Analytics for Relation-driven Problem Environments", that addressed the complexity of and the use of graphical methods for associating and fusing complex entity-relation structures that comprise situation and threat component and holistic states. Relations are the key component of all these ontologies in part because any entity-pair or set may have a multiplicity of relations, some directed, some not, and all can be temporally dynamic; these properties add complexity to the understanding of and the labeling of any entity subnetwork. It is not clear how a priori defined ontologies can address the issue of emergent properties and their dynamics. Relatively little research has occurred in the data fusion community on these topics since that time; see (Llinas, 2021).

Epistemology: Epistemic Pluralism

Epistemic pluralism emphasizes the value of a) a multitude of perspectives, approaches, methods, and types of knowledge (Kellert et al., 2006), and b) "contextualism", which highlights the exploitation of the varied contexts in which an IA problem evolves, e.g., (Stegenga, 2014) and (Veit, 2020) for a taxonomy of various model structures that could be used in pluralistic approaches. We note that the data fusion community has had an emphasis on including the effects of contextual parameters and conditions now for some time, and in some sense therefore having an epistemic pluralism attitude; an entire book addresses the issue of accounting for and exploiting contextual information and effects in (Snidaro et al., 2016). To give some real-world meaning to these ideas, there have been a number of papers addressing the epistemically-based controversies of the Covid-19 period. One of the key factors driving the epistemic battles of that period were the contested prioritization of different types of scientific knowledge. Examples of these different types of scientific knowledge include results and claims that differ in their empirical sources and underlying study designs, that had a wide variety of (different) features. Various authors in the medical field (e.g., (Canali and Lohse, 2024)) have put forward the need for the medical community to instead move away from prioritization of particular approaches and beyond unproductive epistemic battles when dealing with fundamentally complex problems. They call for the employment (and open-mindedness) of *epistemic pluralism*, which emphasizes the utility of a multitude of perspectives, approaches, methods, and types of knowledge.

Another reason to look to epistemology for guidance is related to the idea of the justification of beliefs in formed hypotheses. That is, what standards can be used to determine when we can be justified in believing something to be true? Another way to think about this question is to ask how it is that a *sufficient* standard for truth (an epistemically-defendable standard), rather than an absolute one, can be asserted. From (Whitesmith, 2022): "Epistemic justification matters to intelligence analysis because it is the link between the ideas of truth and evidence", and "Intelligence analysis is fundamentally an exercise in epistemic justification."

To our understanding, the IA community has not seriously examined the potential of epistemic pluralism or any serious consideration of epistemology in any serious way. In the intelligence community of practitioners with few theorists, the predominant approaches assume consistency of patterns and models, regularities, and a conventional reductionist approach. Such approaches assume that knowledge is based on static and well-established laws, rules, and behaviors. Marrin also makes a similar appeal for the IA community on the issues related to intelligence analysis becoming a profession versus a tradecraft in his book (Marrin, 2012) where he considers a wide variety of factors for the failure of the development of an intelligence theory, and a mixed community of practitioners and theorists.

Epistemological Communities

From the views developed in this paper, we believe, as in (Herbert, 2006), that "the purpose of intelligence analysis is the wise management of epistemic complexity…. the intelligence analyst is above all an explainer of epistemic situations". And, following (Hulnick, 2006), that "the intelligence community needs to develop a twenty-first century analytic culture that differs from the conventional intuitive analysis of the past." These views relate to the idea that Intelligence analysis is a knowledge-building activity, and improved analysis requires an understanding of epistemology or the theory of the origin of and the nature of relevant knowledge.

SUMMARY

This paper has addressed complicated topics, ranging from credition and the notions of belief to those of consciousness, uncertainty, and the underlying complexities of real-world problems and processes that include emergent and other properties. These complex and subtle factors make the discussions about and modeling of human-AI process communication necessarily complex as well, and to develop capabilities for believable, effective, reliable human-AI communication imputes the need to address all these factors if a workable and trustworthy capability is to be realized. To build AI capabilities helpful for analyzing complex adversarially-based problems also demands developing a view of the complexity of that problem space. We suggest that the reality of those problems will initially require forming an epistemological stance about the analytical knowledge-based approaches required.

REFERENCES

Ahn, A. C., Tewari, M., Poon, C. S., & Phillips, R. S. (2006). The limits of reductionism in medicine: Could systems biology offer an alternative? PLoS medicine, 3(6), e208.

Alangari, N., El Bachir Menai, M., Mathkour, H., & Almosallam, I., Exploring evaluation methods for interpretable machine learning: A survey. *Information*, *14*(8), 469, (2023).

Bērziņš, J. (2020). The theory and practice of new generation warfare: The case of Ukraine and Syria. The Journal of Slavic Military Studies, 33(3), 355–380.

Canali, S., & Lohse, S. (2024). How to move beyond epistemic battles: Pluralism and contextualism at the science-society interface. Humanities and Social Sciences Communications, 11(1), 1–5.

Duvenage, M. A. (2010). Intelligence analysis in the knowledge age: An analysis of the challenges facing the practice of intelligence analysis (Doctoral dissertation, Stellenbosch: University of Stellenbosch).

Esmaeilzadeh, H., & Vaezi, R. (2021). Conscious AI. *arXiv preprint arXiv:2105.07879*.

Guo, Z., Wan, Z., Zhang, Q., Zhao, X., Zhang, Q., Kaplan, L. M.,... & Cho, J. H. (2023). A survey on uncertainty reasoning and quantification in belief theory and its application to deep learning. *Information Fusion*, 101987.

Herbert, M. (2006). The intelligence analyst as epistemologist. International Journal of Intelligence and Counter Intelligence, 19(4), 666–684.

Hulnick, Arthur S. 2006. "What's Wrong with the Intelligence Cycle." 2006.

Intelligence Community Directive 203: https://www.odni.gov/files/documents/ICD/ICD-203_TA_Analytic_Standards_21_Dec_2022.pdf

Kellert, S. H., Longino, H. E., & Waters, C. K. (Eds.). (2006). Scientific pluralism (Vol. 19). U of Minnesota Press.

Landgrebe, J., & Smith, B. (2022). *Why machines will never rule the world: Artificial intelligence without fear*. Taylor & Francis.

Llinas, J. (2012, September). Framing and Defining New Fusion Strategies and Advanced Analytics for Relation-driven Problem Environments. In Natl Symp on Sensor and Data Fusion, Washington DC.

Llinas, J., Thoughts on Research Imperatives in Data Fusion. In AIAA Scitech 2021 Forum (p. 0914).

Lumbreras, S. (2022). The synergies between understanding belief formation and artificial intelligence. *Frontiers in Psychology*, *13*, 868903.

Marrin, S. (2012). *Improving intelligence analysis: Bridging the gap between scholarship and practice*. Routledge.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16, 1–85.

Senge, R., Bösner, S., Dembczyński, K., Haasenritter, J., Hirsch, O., Donner-Banzhoff, N., & Hüllermeier, E. (2014). Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, *255*, 16–29.

Smith, B. (2012). Ontology. In The furniture of the world (pp. 47–68). Brill.

Snidaro, L., Garcia-Herrera, J., Llinas, J., & Blasch, E. (2016). Context-enhanced information fusion. *Boosting Real-World Performance with Domain Knowledge*.

Solum, L. B. (2014). Artificial meaning. *Wash. L. Rev.*, *89*, 69.

Stegenga, J. (2014) Down with the Hierarchies. Topoi 33(2):313–322. https://doi.org/10.1007/s11245-013-9189-4

Veit, W. (2020). Model pluralism. Philosophy of the Social Sciences, 50(2), 91–114.

Vestrucci, A., Lumbreras, S., & Oviedo, L. (2021). Can AI Help Us to Understand Belief? Sources, Advances, Limits, and Future Directions.

Whitesmith, M. (2022). Justified true belief theory for intelligence analysis. Intelligence and National Security, 37(6), 835–849, doi: 10.1080/02684527.2019.1710806 https://www.researchgate.net/publication/245493621.