ESO-5W1H Framework: Ontological model for SITL paradigm

Shubham Rathi[1] and Aniket Alam[2]

[1] IIIT Hyderabad, Hyderabad, India
shubham.rathi@research.iiit.ac.in
[2] IIIT Hyderabad, Hyderabad, India
aniket.alam@iiit.ac.in

Abstract. The HITL paradigm has been extended as SITL (Society-In-The-Loop) to account for the broader role of AI in the society and vice versa. To open up these otherwise opaque systems and their nexus of interactions with humans, there is a need to make tools to program and debug the algorithmic social contract, a pact between various human stakeholders, mediated by machines. In this paper, we propose one such tool, the ESO-5W1H framework to adjudge the role of humans and machines in their respective interactions and to structure the underlying decision making process such that accountability and liability for each system action-interactions can be brought to the fore. We discuss the working of this conceptual framework in the context of three use cases: the Self-driving car, an AI-based jury, and Neural Networks.

Keywords: Human-In-The-Loop · Society-In-The-Loop · 5W1H · ESO · Ontology · Accountability

1 Introduction

Computer Science and Artificial Intelligence (AI) has penetrated into many domains besides Information technology. AI is now beyond a tooling role and is applied for autonomous operations in domains like Banking- where it is used to determine creditworthiness [1], Medicine - in diagnostics [2], Construction - for Township and building planning [4], judiciary - in risk assessment [3] and even recruiting. These all domains have historically been such where human judgment and ethics have always been a decisive factor in the outcome. As a stop-gap arrangement, these AI tools are used by keeping humans in the loop to augment rather than automate the decision process and to bring in more accountability and transparency. These technologies are not purely autonomous as in many cases; humans do the rule framing and training. Hence, there is always a possibility that these systems pick up inadvertent bias from its creators. Thus, The debate around liability and autonomous systems needs to be reframed more precisely to reflect the agentive role of designers and engineers and the new and unique kinds of human action attendant to autonomous systems that would help fade the black-box reputation which AI is infamously earned and bring in more regulation.

This paper builds on the idea of ‘Society-in-the-loop’ (SITL) paradigm [5] that maps the larger societal role in the development of AI technologies. As the impact of AI technologies spills over the society, there is a need to adjudge them in a framework, in an algorithmic social contract [5] that is mediated between the various stakeholders and the machines. To implement SITL, there is a need to form new feedback loops that embed the social, cultural and quantifiable morals into the system for which, there is a ‘need to build new tools to program, debug, and monitor the algorithmic social contract between humans and algorithms’ [5]. Our approach is tool cum framework that aims at formalizing a structure for such tasks.

The remainder of the paper is organized as follows: The next subsection speaks further on the background and the need for such a framework. In Section 2, we introduce the ESO-5W1H framework. Section 3 explains the working of this conceptual framework concerning its applicability in end to end systems (Self-driving car) and also shows the viability of such an approach for state-based systems like Neural Networks. Section 4 concludes the paper with recommendations for future work.

1.1 Background & Motivation

Management Science has numerous frameworks through which transparency and accountability is established in organizations. Notably, the Fishbone analysis [7] and the 5W-1H approach [8] have been applied for root cause analysis in Software Engineering. A similar framework is needed for evaluating the decisions taken by an Artificially Intelligent (AI) agent. This paper intends to give a knowledge engineering based extension to the causal aspects of AI thinking that is currently overlooked and cut off at the machine level. Our framework is a step towards building a more regulated and responsible AI framework that is overarching enough to trace its roots to respective human, non-human agents in the process loop. Example, if a Neural Network is known to discriminate on the creditworthiness of a candidate on gender, it is a hunch that the bias is perhaps in its training data and thus the responsibility of its designer. However, since no causal chain can link this to its cause, it is always a guessing game as to what part of the system needs a tweak. By bringing an ontological perspective to the problem, questions like, “What happens when there is no direct human actor, only a computational agent - responsible?” becomes “How do we locate the network of human/ non-human actors responsible for the actions of computational agents?” [6].

Model Interpretability There is substantial work in interpretable machine learning aimed at trying to gain visibility into the models. This body of research is mostly around eliciting Post hoc interpretations from models and is largely in four categories [19]:

- Text explanations: Since humans understand explanations verbally, one model might be trained to generate predictions, and a separate language model to generate explanations. This approach was demonstrated by Krening et all [20]. McAuley and Leskovec demonstrate the use of text to explain decisions of a Latent Text model [21].

- Visualization: This approach is to generate corresponding visualizations that will help decompose the model. This was demonstrated by Olah, Chris, et al [22] where they work on feature visualizations with an intent of gleaning into its semantic factors. To understand what information is retained at various layers of a neural network, Mahendran and Vedaldi [23] pass an image through a discriminative CNN to generate a representation. They then demonstrate that the original image can be recovered with high fidelity even from reasonably high-level representations (level 6 of an AlexNet) by performing gradient descent on randomly initialized pixels [19]
- Local explanations: While it may be difficult to describe succinctly the full mapping learned by a neural network, some of the literature focuses instead on explaining what a neural network depends on locally [19]. A popular attempt at local explanations is by Ribeiro et al [24] where they propose a tool, 'LIME' (Local Interpretable Model-Agnostic Explanations) to explain the prediction of any classifier. The intuition behind LIME is that behavior of a model can be learnt by perturbing the input and evaluate the prediction change.
- Explanation by example: This approach is similar to visualizations and uses attention based RNNs to explain by analogy, to report (in addition to predictions) which other examples are most similar with respect to the model as is shown by Olah, Chris, et al [22] and Caruana et al [25]

Need for Ontology Many researchers have made the need of an Ontology implicitly known. Rahwan [5] calls for building new tools to program, debug, and monitor the algorithmic social contract between humans and algorithms. One such tool is Ontology. Ontology is also useful where Bieger et all [9] seek white box evaluation methods for AI that internal functioning and system behavior could be understood in terms of the what, why and how of the outputted result.

With the increasing infiltration of autonomous products into the shared public space, there is a need to have an ethical, moral and a social basis to its activities and existential nature best expressed in an ontological form. As discussed in the previous section, there do exists evaluation metrics for the fidelity of a model, but there is a very loose translation of these metrics to terms comprehensible by a non-domain expert. It is here at ontology can complement and value add the efforts ongoing in model interpretability.

Society in the Loop: This paper is intended to be an extension to the Society in the loop framework proposed by Iyad Rahwan [5]. In his paper, Rahwan discusses the current problems in AI, notably:

- Black Box notion: AI and its underlying technology and outcome is very intricate and almost a black box to its stakeholders which erodes the notion of accountability.

- Filter Bubbles: There is a concern that people succumb to living in filter bubbles created by news recommendation algorithms and user based profile targeting techniques.
- Design Bias: Data-driven decision support system can perpetuate injustice by picking up inadvertent human biases from the training data.

There have been different solutions proposed for the above problems, most of which are policy based. The United States White House National Science and Technology Council Committee on Technology [10] released recommendations ranging from eliminating bias from data to regulating autonomous vehicles to introducing ethical training in computer science curriculum. The European Union, which has enacted many personal data privacy regulations, has proposed granting robots legal status to hold them accountable, and to produce a code of ethical conduct for their design [11]. IEEE has produced a charter on 'Ethically Aligned Design' [12], a crowd-sourced global treatise regarding the ethics of Autonomous and Intelligent Systems. Rahwan has gone a step further and has tied this policy into a formal Society-In-The-Loop paradigm where he defines SITL crisply as SITL = HITL + Social Contract.

SITL = HITL + Social Contract: The HITL idea only serves a narrow, well-defined function having very specific use cases: Labelling Data, Interactive Machine Learning [13], Systemized applications as in a crisis counseling system [14]. Rahwan argues that for a system which has a more societal impact and implication, like an AI algorithm that controls many self-driving cars or a news filtering algorithm influencing political beliefs or algorithms that determine creditworthiness thereby affecting the allocation of resources - the SITL paradigm comes to picture. He states, 'While HITL AI is about embedding the judgment of individual humans or groups in the optimization of AI systems with narrow impact, SITL is about embedding the values of society, as a whole, in the algorithmic governance of societal outcomes that have broad implications.' These societal values are the 'Social Contracts' that an individual implicitly gets in with society. Thus in the SITL domain, there must be a general agreement on the accepted tradeoffs between the different values that AI systems can strive for and have a demarcation on which stakeholders reap what. Rahwan's framework is particularly signification as it brings about a formal framework for implementing the policy considerations proposed earlier. Our framework aims to fill the gaps that the SITL idea surfaces:

- Need for new metrics and methods to evaluate AI behavior against quantifiable human values.
- Bridge the cultural divide between engineering and humanities by having a common vocabulary to articulate policy expectations to engineers and designers. It is difficult to quantify the behavior of systems such that ethicists and legal experts easily understand them.
- Quantifying negative externalities (cost incurred by third parties) is difficult due to long and opaque causal chains in the machine process. Reading the

source code of any modern machine learning algorithm tells little about its behavior as discrimination often emerges through the interaction of data and the algorithm.

– Negotiating the tradeoffs is difficult because of many interacting agencies.
– Ensure that algorithms are performing as expected.

2 ESO-5W1H Framework

The ESO-5W1H model is based on a two-tier homogeneous ontology. The upper ontology is the Event and Implied Situation Ontology (ESO) [3] which is the substratum for the 5W1H model. The 5W1H model is based on the 5W1H maxim: Why, What, When, Where, Who and How which is widely used in management studies for cause-effect analysis. The system at this stage is only a conceptual schema, and the focus of this paper is to highlight the framework and the use cases only. The task of exhaustively listing classes, relationships and the mappings between 5W1H and ESO are beyond the scope of this paper's discussion but relevant nonetheless.

2.1 ESO Ontology

ESO reuses and maps across existing resources such as WordNet, SUMO, and FrameNet and is designed to facilitate implicit reasoning. Following best practices in Semantic Web technologies, ESO reuses parts of two existing vocabularies: there are mappings from ESO to Framenet on class and role level and mappings to SUMO on class level [16]. ESO models the implications before, after and during the event including the role of the involved entities. Example a statement like: 'Apple hired Steve Jobs to save the company' could be modeled as:
- Before: Steve *notEmployedAt* Apple
- After: Steve *EmployedAt* Apple
-           Steve *hasTask* save the company
-           Steve *isEmployed* true

We do not get into the details of the ESO classes and relationships as the contribution of the paper is the symbiosis of ESO with 5W1H. However, a detailed documentation on ESO its classes, attributed and relation can be found in the ESO documentation[4]. Note that the ESO classes, relations are derived from SUMO, so even if there exist classes and relations which are not defined in the ESO documentation, they could be sourced from SUMO and the ontology could be held viable for a variety of use cases.

[3] http://www.newsreader-project.eu/results/event-and-situation-ontology/

[4] `https://github.com/newsreader/eso-and-ceo/blob/master/ESO_Documentation.pdf`

2.2 5W1H Ontology

Our 5W1H ontology is partly based on the CA5W1HOnto Ontology [15] - it has the same top level classes but different entities and relations. The relations and entities are derived from SUMO and ESO to ensure there is maximum overlapping with the ESO ontology. The broad structure of top classes is depicted in figure 1:

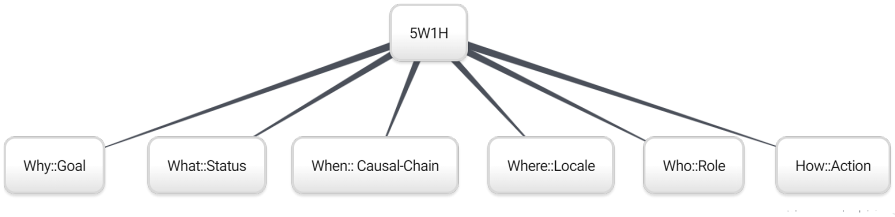


Fig. 1. 5W1H top level classes

As evident, the 5W1H is the feature of this framework that brings in granular level accountability.

2.3 ESO-5W1H Metamodel

The idea is that ESO ontology will formulate the worldview representation that will be passed down to the 5W1H model. Though the original ESO ontology was tested only on textual data, there is no hindrance to assume and generalize that ESO can be coupled to work with non-textual data too. E.g., In a computer vision system that populates the ESO model based on whatever it captures. The ESO model is capturing the sequences of states and their changes over time. This is a noisy representation as the system is capturing all the details, even the ones which are not relevant to the scene. A curation on this data is necessary. For this activity, we propose to use an Actor-Network Engine that will assemble the 5W1H network from the ESO. There is no special emphasis on the use of Actor-Network theory concepts except for borrowing its vocabulary and network formation ideas. There could be a parallel discussion on network formulation alternatives.

The Actor-Network engine works via a process known as Translation in the Actor-Network Theory. Translation is further simplified into 4 discrete steps:

- Problematisation: Defining the problem and the primary actor
- Interessement: during which the primary actor(s) recruit other actors to assume roles in the network
- Enrolment: during which roles are defined, and actors formally accept and take on these roles

- Mobilisation: during which primary actors assume a spokesperson role for passive network actors (agents) and seek to mobilize them to action.

Translation results in the formation of the 5W1H model underneath. At the Problematisation stage, the 'Who::Role' class is exhaustively populated. During the Interessement stage, the 'What::Status' and the 'Where::Locale' features are set up. As the network matures, secondary actors are eliminated and the 'Why::Goal' and the 'When::CausalChains' are decided in the Enrolment Phase. As a final step in the network formation, Possible action strategies are populated in the 'How::Action' class. The resulting 5W1H network is a subset of the original ESO model with the 5W1H classes (Who, When, Where, Why, What, How). The system at this stage is still not ready to act since the social contract has still not validated. Thus, the system makes calls to the Algorithmic Social Contract (ACS) system and negotiates a tradeoff strategy and finally acts on it. The ACS needs manual rules to be embedded in situations of uncertainty and thus extends finally to the HITLs to program these social contract rules. This process is explained in Figure 2.

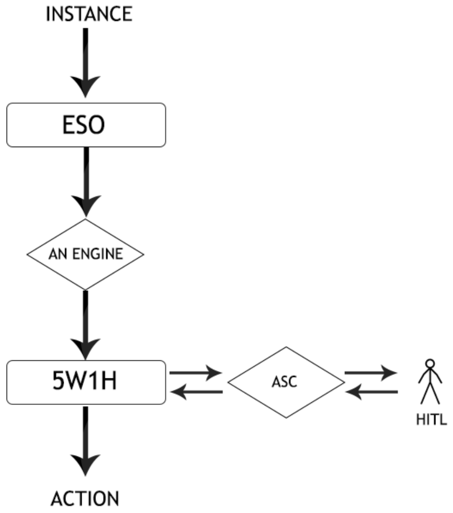


Fig. 2. Pipeline of ESO-5W1H

This pipeline is a systemic way of implementing the SITL paradigm. The system has been kept open-ended at many touchpoints to account for the different architectures that could be hacked together to achieve the same goal, of embedding social contract into the system behavior and to have implementation level accountability into the system action. A few use cases in the following section should clarify the working of this framework and the interaction between various subsystems.

3 Usecases

In this section, we bring visibility to the working of the pipeline and demonstrate how this framework brings resolution to the black box problem. An ideal application of such an ontology is with the task of translating model interpretability to a lay man. An ontology is a complementary interface that could sit on top of systems like LIME, Eli5 and translate model fidelity with minimum technical jargon. As we shall see in cases below, this ontology can be applied in range of cases where we have to embed societal contracts and also sit on top of such systems and complement their.

3.1 Case 1: Embedding the Social Contract

Consider a social contract that is followed on the streets in pedestrian crossings. In most Asian countries, if the pedestrian makes eye contact with a driver, the right of way is given to the car. In most western countries, if Pedestrian makes eye contact with the driver, it implies that the driver is to give the right of way to the pedestrian. Thus, even the social contracts are not uniform and may vary depending on the cultural and geographical context. A case for such a scenario may be made as follows:

– Instance: “The eye tracker on the car camera reports eye contact with a Pedestrian”
– ESO: The system spawns an ESO representation of the instance:
  • Pre Situation:
    - Pedestrian *notAtPlace* Crossing
    - Signboard *AtPlace* Crossing
    - Car *inState* Motion
    - CarCameraSensor *hasAttribute* No-contact

  • During Situation:
    - Pedestrian *AtPlace* Crossing
    - Signboard *AtPlace* Crossing
    - Car *inState* Motion
    - CarCameraSensor *hasAttribute* Made-contact

– AN-Engine: The change triggers the AN-Engine which initiates the 5W1H network formulation (Translation) which distills the relevant details for the system to process.
  • Problematisation:
    - *Problem Definition:* Action to Eye contact
    - *Who::Role:* Car, Pedestrian, Signboard, Road

- Interessement:
  - *What::Status:* Car *inMotion* True, Pedestrian *inMotion* False, Powerbreak *isActive* true, Powerbreak *inFunction* false, GeoSensor *isDamaged* false, Car *hasAttribute* Speed, Speed *hasValue* 30-mph
  - *Where:: Location* Car *atPlace* 4th Street, Pedestrian *atPlace* 4th Street Crossing

- Enrolment: Eliminating secondary actors - *Who::Role:* Car, Pedestrian. *Why::Goal* - Policy for right of way
- Mobilisation: *How::Action* = (Stop Car, Slow Car, Continue pace, Switch to manual, Alert Driver, Continue in Automatic), Fixing causal chains for *When::CausalChains*

– 5W1H: The above process assembles in a 5W1H network as follows:
  - *Why::Goal* = Policy for right of way
  - *What::Status=* Car *inMotion* True, Pedestrian *inMotion* False, Powerbreak *isActive* true, Powerbreak *inFunction* false, GeoSensor *isDamaged* false, Car *hasAttribute* Speed, Speed *hasValue* 30-mph
  - *When::CausalChain* = CausalChainN(Car-In-Motion)→CausalChainN+1(Pedestrian) →CurrentFrame.
  - *Where:: Locale=* Car *atPlace* 4th Street, Pedestrian *atPlace* 4th Street Crossing
  - *Who:: Role=* Car, Pedestrian
  - *How::Action=* (Stop Car, Slow Car, Continue pace, Switch to manual, Alert Driver, Continue in Automatic)

– ASC: The 5W1H model makes calls to the Algorithmic Social Contrast subsystem which negotiates a tradeoff on the action strategy. If the car was operating in Asian context, the ASC system would yield a feedback as: *Slow Car→Continue Pace →Alert Driver.*

  However, if the car was in a western context the feedback would be different: *Slow Car→Stop Car →Alert Driver.*

– Action: Assuming the car was in a Asian context, the action taken would be: *Slow Car→Continue Pace →Alert Driver.*

3.2 Case 2: Isolating Bias

Assume a hypothetical situation wherein an AI system is part of a jury and has passed a verdict against the John over Diana even when the facts were inconclusive. Without the 5W1H model, it is not possible to gain visibility into the decision process. If the 5W1H query was possible, the following log could have been been found:

– Query: Why(5W1H - John guilty)

The system does a traceroot call to the last frame where the network was Mobilizing that John was guilty. The system state at this stage:

– What::Status= Fact1(Inconclusive), Fact2(Inconclusive), Fact3(Inconclusive)

- When::CausalChain = CausalChain1 →CausalChain2 →CausalChain3 ..
- Where::Locale= XXYY
- Why::Goal= Evaluation of facts
- Who::Role=John (PersonID1232), Diana(PersonID2211) ..
- How::Action= WeightedAverage(Facts, Legal Precedents)

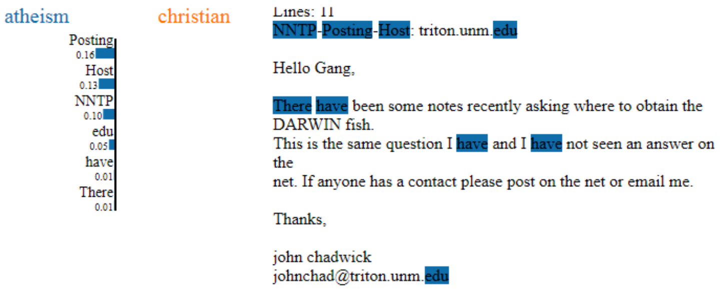
Looking at this frame, Its still not conclusive why AI reached the decision but there is a clear picture that the facts were inconclusive even for the system which leaves only the legal precedents to investigate.

- Query: What(Legal Precedent)

This query returns a dataset of legal precedents of similar charges. In most of these cases, since men were having a higher crime rate than women, the system evaluated this decision on this statistical truth and thus convicted John. This decision by the AI is a blunder since our legal regimes prohibit discrimination on the basis of sex. It is a social contract in the society that justice shall be equal irrespective of caste, creed, sex, and color. Thus, even without embedding a bias in the system, the system picked up bias from the data. Had an algorithmic social contract system existed in this system, the option of relying on legal precedents on making a verdict would have been eliminated (because it would be embedded in the social contract to not factor in discrimination on the basis of gender) and the system would not have made a faulty conviction.

3.3 Case 3: With state based systems

The proposed framework is not only valid for end to end systems as discussed above but is also for state based systems like Neural Networks, decision trees and other classifiers. This is possible if an ontology can be coupled with tools like LIME, Eli5 and attention based RNN models. For the sake of example, we can consider a case from LIME. When applied on the 20 newsgroup dataset, LIME reveals an interesting observation about the classification. When making a classification between Christianity vs atheism, the classifier relies on the metadata from the email header as a decisive classification criteria.



Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the
DARWIN fish.
This is the same question I have and I have not seen an answer on
the
net. If anyone has a contact please post on the net or email me.

Thanks,

john chadwick
johnchad@triton.unm.edu

Fig. 3. Classification metrics as revealed by LIME

This classification metric is wrong as it relies on a non-universal feature. To generate a pseudo natural language explanation for this outcome, a system like LIME will have to be interfaced with an ontology so that higher order logics behind the classifier behavior could be traced. With the GDPR guidelines making 'right to explanation' [26] into policy - Recommender and classification systems will have a handy use of ontologies to casually explain its blackbox behavior. It is to be noted here that the algorithmic contract here is to rely on on universal features. Since the email header is not a universal feature, the engine would flag this classification as faulty.

4 Conclusion & Future Work

In this paper, we propose a generic framework to implement the SITL paradigm at a systemic level. The framework proposed is abstract and is intended spur discussions on a computer science perspective and its corresponding implementations. Using various use cases, we demonstrate how this pipeline is proposed to work and how it absolves the various black box problems surfaced by Rahwan [5]. This paper also demonstrates the need and benefit of ontological integration into the SITL thought process.

Future work on this paper will be to solidify the subsystem components discussed in the paper (AN Engine, ASC). Besides the ontology, the AN Engine and the ASC subsystem of the system still needs architectural brainstorming. The rules and the processes required for such a robust system will have to be carefully drafted. The SUMO classes will have to be evaluated for cross-domain applicability. The classes have to be generic enough to account for any event and situation. As demonstrated in case 3, the most obvious and usable development in the area is to build a system around the integration of ontologies with model interpretation frameworks.

A significant challenge in the system will be to generate consensus and agreement on which social contracts are acceptable and which are not. The elimination of a few social contracts will generate less accurate outputs, but in the interest of a fair AI entity, that tradeoff will have to be met. Example: In use case 2, eliminating the statistical truths about crime rate will render less accurate decisions but since the AI cannot be given to judge when it is appropriate to factor it in and when to not, it is necessary that the data about gender ratio in crime rate be purged altogether. A uniform agreement on many such cases will have to be debated and agreed.

References

1. Islam, Md, Lin Zhou, and Fei Li. "Application of artificial intelligence (artificial neural network) to assess credit risk: a predictive model for credit card scoring." (2009).
2. Kamruzzaman, S. M., et al. "Medical diagnosis using neural network." arXiv preprint arXiv:1009.4572 (2010).

3. Geraghty, Kate Anya, and Jessica Woodhams. "The predictive validity of risk assessment tools for female offenders: A systematic review." Aggression and violent behavior 21 (2015): 25-38.
4. Yatabe, S. M., and A. G. Fabbri. "Artificial intelligence in the geosciences: a review." Sceinces de la Terre, Ser. Inf., Nancy 27 (1988): 37-67.
5. Rahwan, Iyad. "Society-in-the-loop: programming the algorithmic social contract." Ethics and Information Technology 20.1 (2018): 5-14.
6. Elish, M. C., and Tim Hwang. "Praise the Machine! Punish the Human! The Contradictory History of Accountability in Automated Aviation." (2015).
7. Slaughter, Sandra A., Donald E. Harter, and Mayuram S. Krishnan. "Evaluating the cost of software quality." Communications of the ACM 41.8 (1998): 67-73.
8. Chung, Sam, et al. "Service-oriented reverse reengineering: 5W1H model-driven redocumentation and candidate services identification." Service-Oriented Computing and Applications (SOCA), 2009 IEEE International Conference on. IEEE, 2009.
9. Bieger, Jordi, et al. "Evaluation of general-purpose artificial intelligence: why, what & how." Evaluating General-Purpose AI (2016).
10. National Science and Technology Council Committee on Technology. (2016). Preparing for the future of artificial intelligence. Technical report, Executive Office of the President.
11. Delvaux, M. "DRAFT REPORT with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103 (INL))." Committee on Legal Affairs. European Parliament, PR1095387EN. doc, PE582. 443v01-00 (2016).
12. IEEE Global Initiative. "Ethically Aligned Design." IEEE Standards v1 (2016).
13. Cuzzillo, T. "Real-world active learning: Applications and strategies for human-in-the-loop machine learning." (2015).
14. Dinakar, Karthik, et al. "Mixed-initiative real-time topic modeling & visualization for crisis counseling." Proceedings of the 20th international conference on intelligent user interfaces. ACM, 2015.
15. Kim, Jeong-Dong, Jiseong Son, and Doo-Kwon Baik. "CA 5W1H onto: ontological context-aware model based on 5W1H." International Journal of Distributed Sensor Networks 8.3 (2012): 247346.
16. Segers, Roxane, et al. "Eso: A frame based ontology for events and implied situations." In Proceedings of Maplex2015. Naushad. 2015.
17. Uchida, Yusuke, et al. "Embedding watermarks into deep neural networks." Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. ACM, 2017.
18. Johnson-Laird, Andy. "Neural networks: The next intellectual property nightmare?." COMP. LAWYER. 7.3 (1990): 7-16.
19. Lipton, Zachary C. "The mythos of model interpretability." arXiv preprint arXiv:1606.03490 (2016).
20. Krening, Samantha, et al. "Learning from explanations using sentiment and advice in RL." IEEE Transactions on Cognitive and Developmental Systems 9.1 (2017): 44-55.
21. McAuley, Julian, and Jure Leskovec. "Hidden factors and hidden topics: understanding rating dimensions with review text." Proceedings of the 7th ACM conference on Recommender systems. ACM, 2013.
22. Olah, Chris, et al. "The building blocks of interpretability." Distill 3.3 (2018): e10.
23. Mahendran, Aravindh, and Andrea Vedaldi. "Understanding deep image representations by inverting them." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

24. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. ”Why should i trust you?: Explaining the predictions of any classifier.” Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.
25. Caruana, Rich, et al. ”Case-based explanation of non-case-based learning methods.” Proceedings of the AMIA Symposium. American Medical Informatics Association, 1999.
26. General Data Protection Regulation (GDPR), 2016/679 (Official Journal of the European Union 2016