A Model for Artificial Conscience to Control Artificial Intelligence

Davinder Kaur, Suleyman Uslu, and Arjan Durresi

Abstract We propose *"Artificial Conscience - Control Module"* framework to control the AI systems and to make them adaptable based on the user requirements. AI users can have different needs from the same AI system, and these systems must adjust their output based on these requirements. The proposed framework enables users to provide context to the AI system by assigning weights to different evaluation metrics. Based on these weights, AI metrics-agents negotiate with each other using our trust engine to output a solution with maximum *"Artificial Feeling."* This framework can be easily implemented in any AI system where multiple metrics are involved. We have illustrated the proposed framework using an AI system for classifying people based on income.

1 Introduction

Artificial Intelligence(AI) systems have transformed our lives. Nowadays, almost every task is either guided or done by algorithms. The rapid development and growing use of these systems have raised many concerns. These systems have become complex and do not always yield safe and reliable results. Taking proper measures to design, develop, test and oversee these systems becomes very important. Different researchers have proposed various ways to accomplish this.

Davinder Kaur
Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA e-mail: davikaur@iu.edu

Suleyman Uslu
Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA e-mail: suslu@iu.edu

Arjan Durresi
Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA e-mail: adurresi@iupui.edu

Some government and private research organizations have proposed different ethical guidelines and frameworks to make AI safe, reliable, and trustworthy. One of the primary requirements proposed by most of these frameworks is the involvement of the human agency to control AI [14]. Researchers have proposed the concept of bounded optimality [17, 19] to control AI. Another line of researchers has proposed the concept of artificial conscience, which deals with replicating some aspects of the human consciousness in machines [4]. Baasr, a cognitive neuroscientist [1] proposed the Global Workspace Theory of brain consciousness using a theatre analogy. Solms [21, 20] proposed that consciousness is related to feelings. Blum proposed the theory of consciousness through a computer science perspective [3], in which different processors compete with each other to get their information broadcasted to other processors.

In this paper, we have combined both lines of research on controlling AI and artificial conscience. We developed a model of an artificial conscience-control module, which can be used to make AI systems adapt according to the users' requirements using the concept of "Artificial Feeling." This control module provides controllability of AI system decisions, making them safe, reliable, and trustworthy and hence increasing their acceptance in society. This paper is organized as follows. Section 2 presents the background and related work in the field of AI, artificial conscience, and trust. Section 3 describes our proposed framework, "Artificial Conscience Control Module." Section 4 illustrates our framework using an AI system for classifying people based on their income, and in Section 5, we conclude our paper.

2 Background and Related Work

This section presents background and related work for the need to control AI, artificial conscience, and the role of trust.

2.1 Need to Control AI

The wide adoption of AI systems does not imply that they are always safe and reliable [9]. It becomes essential to control and oversee these systems to prevent any harm caused by them to the users or society. Different government and private research organizations have proposed various guidelines and frameworks to make them safe, reliable, and trustworthy. One of the main requirements presented by these agencies is the involvement or control of humans in AI decision-making [14]. Different researchers have proposed various ways to involve humans in the AI lifecycle. European Union (EU) [6] suggested the involvement of humans in three phases: designing, developing, and overseeing. Other researchers have proposed the involvement of humans based on the risk associated with using AI. International Organization for Standardization (ISO) [7] also suggested the involvement of humans

by integrating control points in the AI life cycle to increase the trust and adoption of AI systems. The proposed frameworks by all these influential organizations show the importance of human involvement in controlling AI.

2.2 Artificial Conscience and Controlling AI

Artificial conscience, also known as machine conscience, is a way to implement some aspects of human cognition that comes from the phenomenon of consciousness [4]. Different researchers have proposed other goals that can be achieved by artificial conscience. Some of them are autonomy, resilience, self-motivation, and information integration. To achieve these goals, there is a need to design conscious machines that can replicate some features of the conscious experience.

To replicate some features of the conscious experience, it's imperative to understand what consciousness is. Baars, a cognitive neuroscientist [1], proposed the Global Workspace Theory (GWT) of the brain and explained consciousness through the theater analogy as the activity of actors in a play performing on the stage of Working Memory, their performance under observation by a vast audience of unconscious processors sitting in the dark. Another set of researchers proposed the theory of consciousness through the perspective of theoretical computer science known as Conscious Turing Machine (CTM) or Conscious AI [3]. This theory is influenced by Alan Turing's powerful model of computation known as the Turing Machine and by the global theory of consciousness GWT. Based on this theory, different processors compete with each other to get their information on the stage/short-term memory so that it can be broadcasted to other processors. Solms [21, 20] proposed that consciousness is endogenous and is related to feelings. However, some other researchers argued that the classic notion of rationality is unattainable for real agents [18]. They proposed the concept of bounded optimality [17, 19], which deals with optimizing not the actions taken, but the algorithm used to select the action. These types of agents trade off between efficiency and error. All the work by different researchers deals with understanding the human conscience and how some aspects can be implemented in AI.

2.3 Role of Trust

Trust is a complex phenomenon and is a context-dependent concept. Different disciplines define trust differently. In general, trust is defined as " the confidence one entity has in another entity that it will behave as anticipated" [7]. Trust information is highly influential in decision-making when multiple entities are involved [22]. Different researchers have proposed various ways to calculate and manage trust information. Ruan et al. [16] proposed a trust management framework to quantify trust between entities based on the measurement theory. Because of the flexibility of this

framework, it has been used in various decision-making applications like healthcare [10, 11], social networks [8, 13], crime detection [12], and the food-energy sector [23, 24, 25, 26, 27, 28]. The use of trust in all these applications validates its potential to help capture negotiations between different metrics of AI system.

3 Trust Based Artificial Conscience - A Control Module for AI Systems

This section introduces the trust-based artificial conscience control module to control the AI systems based on the user's needs. This control module assumes that different users can have other requirements from the AI system. Based on their requirement, each user assigns weights to the evaluation metrics of the AI system. Based on these weights, these metrics negotiate with each other over a list of solutions to output the agreed solution. The negotiation between the metrics is controlled by the trust and trust sensitivity of the metrics. Section 3.1 explains the trust engine, Section 3.2 explains the concept of trust pressure and sensitivity, and Section 3.3 describes the framework of the control module.

3.1 Trust Engine

Our trust engine [16] calculates the trust between the entities based on their past interactions. In a decision-making problem, agents interact with each other several times, proposing and rating each other's solutions. The rating provided by an agent to another agent's solution measures the agent's impression of the other agent. The ratings are considered to be between [0,1], where 0 is the lowest rating and 1 is the highest rating. The impression of agent A^X toward agent A^Y, denoted by $m^{X:Y}$, is the mean of the ratings, where $r^{X:Y}$ is the rating of A^X to A^Y given over N number of rounds. Another essential component of trust is confidence, denoted by c. It is used to capture the consistency of the impression. Confidence is inversely related to the standard error of the mean. The formula for impression and confidence is given in Equation 1.

$$m^{X:Y} = \frac{\sum_{i=1}^{N} r_i^{X:Y}}{N} \quad \text{and} \quad c = 1 - 2 * \sqrt{\frac{\sum_{i=1}^{N} (m^{X:Y} - r_i^{X:Y})^2}{N * (N-1)}} \tag{1}$$

The value of the impression and confidence is used to calculate the trust between two agents. This trust framework can also be utilized when two agents are not directly connected using the trust propagation and aggregation methods [16].

3.2 Trust Pressure and Trust Sensitivity

Trust pressure and sensitivity are based on the concept of social psychology, which studies the influence of others on the behavior of individuals [27]. Our framework uses trust pressure and sensitivity to capture agents' changing behavior. The source of the trust pressure, denoted as P, is the difference between the target trust level of the agent, T_{target}, and the agent's trust level from other agents, $T_{current}$. How much the trust pressure affects the agent's behavior, namely effective trust pressure, P_e, depends upon his trust sensitivity, S_T as shown in Equation 2. If the trust sensitivity of an agent is high, it will alter his behavior much faster than an agent with less trust sensitivity. In our framework, trust sensitivity is introduced using the weights of the metrics assigned by the users. These weights are translated into trust sensitivities. Both of them are inversely proportional to each other.

$$P = T_{target} - T_{current} \quad , \quad P_e = P \times S_T \tag{2}$$

3.3 Artificial Conscience - Control Module

We have proposed an artificial conscience control module to control the AI systems based on the user requirements. Users can have different expectations from the AI system based on their needs. The involvement of humans provides meaning to the working of the AI system, which is easily missed by the algorithms alone. For our framework, we assume that there is a decision-making task that needs to be performed by the AI system. Different machine-learning algorithms are deployed for that task, and multiple metrics are used to evaluate the solution. Users assign weights to these metrics based on their requirements to provide context and meaning to the decision-making task. The users' weights are used to calculate the trust sensitivities for the metrics. These metrics, which we call "agents," negotiate with each other based on the ratings/trust they get from other agents and their trust sensitivities to calculate "Artificial feeling" (AF) as a weighted average among agents. This framework consists of the following steps:

- At the beginning of the negotiation, each agent is given the goal of achieving the best option they can get among the pre-computed solutions. Different agents can have different importance on the parameters for their goals, leading to negotiation.
- Agents can negotiate for n number of rounds to reach a solution. n is application or user dependent. In the first round, each agent proposes the best solution based on their needs.
- If the agent receives an adequate level of trust which reflects the other agents' approval, it does not need to alter his solution in the subsequent rounds. Otherwise, it selects the next best solution, which would bring in more trust at the expense of a decline in benefits.

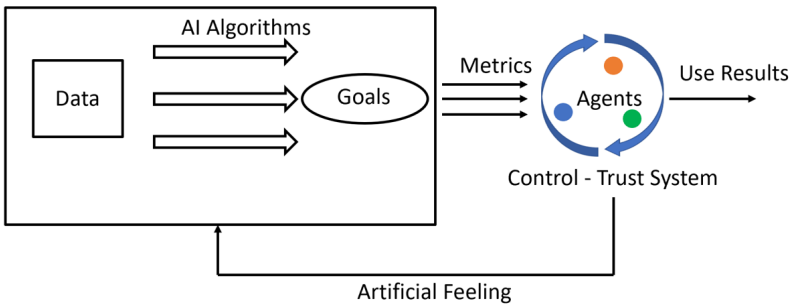


Fig. 1 Artificial Conscience - Control Module to Control AI

- The solutions are compared in terms of the benefit they provide using a distance metric. The distance between a proposed solution p and a goal g is calculated using the euclidean distance formula as shown in Equation 3 where i is a parameter of a solution among d selected parameters for distance calculation.

$$dist(p,g) = \sqrt{\sum_{i=1}^{d} (p^i - g^i)^2} \tag{3}$$

- In each round, agents rate other agents' solutions based on their distance. For example, assume that there are two agents, namely A_A and A_B. If agent A_A proposes a solution closer to its goal but far away from the agent A_B's goal, then the agent A_A receives a low rating from the agent A_B. Based on the rating provided by the other agents, the trust of the agent is calculated using the framework described in Section 3.1.
- In short, each agent tries to minimize the distance from their proposed solution to their goal, considering the trust they receive depending on how much they are sensitive to trust. Their trust is determined by the distance of their proposed solution to other agents' goals. The amount of sacrifice is governed by trust sensitivity. The higher the trust sensitivity, the more the agent sacrifices to raise its rating.
- After n rounds, for each solution, the artificial feeling (AF) is calculated as the weighted average among agents. AF is used for comparing and selecting solutions of different AI algorithms.

This framework enables the control of the AI system based on the user's needs. Figure 1 shows the block diagram of our framework.

4 Implementation

This section describes the dataset used for the experiment, its implementation, and the results using our framework.

4.1 Data

We have performed our experiments on the real-world dataset to consider the accuracy and fairness aspects of the AI algorithm decision-making. US adult income dataset [5] is used for our experiments. The dataset contains 32,561 instances and 14 attributes. It includes two sensitive attributes: race and sex. It is pre-split into training and testing sets for the machine learning prediction task of predicting whether an individual makes more or less than $\$50,000$ per year. For the dataset, oversampling is performed on the minority target variable, and the categorical attributes are converted into numerical vectors using one-hot encoding to be used for training. In our study, we have considered "White" and "Male" as the privileged class for race and sex attributes.

4.2 Experimentation Setup and Results

In our study, the decision-making task that needs to be accomplished by the AI system is to classify the people according to their income. Two supervised machine learning models, namely support vector machine (SVM) and logistic regression (LR), are used to build the prediction model. A post-processing fairness technique, Calibrated Equalized-Odds, is used. This technique optimally adjusts the learned predictors to remove discrimination based on the equalized odds objective [15]. In our experiment, we used two types of metrics. One is the performance metric, i.e., accuracy (A), to quantify how well the system predicts true labels. Another metric type is the fairness metric to evaluate the system's fairness, which is how unbiased the system predictions are concerning sensitive attributes. The three fairness metrics we use are Equal Opportunity Difference (F1), Disparate Impact (F2), and Average Odds Difference (F3) [2]. All these metrics are used to evaluate and compare different algorithms.

We have computed these metrics for sensitive attributes at 25 evenly distributed classification thresholds between (0.01 - 0.99). We labeled these 25 classification thresholds as solutions. In other words, there will be 25 solutions to choose from at each round of decision-making. The trade-off between the accuracy and fairness metrics for the 25 solutions is presented in Figure 2 for post-processing Support Vector Machine (SVM) and in Figure 3 for post-processing Logistic Regression (LR). The trade-off between the accuracy and fairness metrics at different classification thresholds shows that one solution/classification threshold cannot satisfy all the ac-

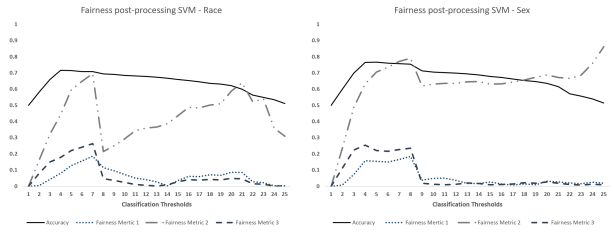


Fig. 2 Accuracy and fairness metrics trade-off for post-processing fairness enhancing SVM on Race and Sex sensitive attributes.

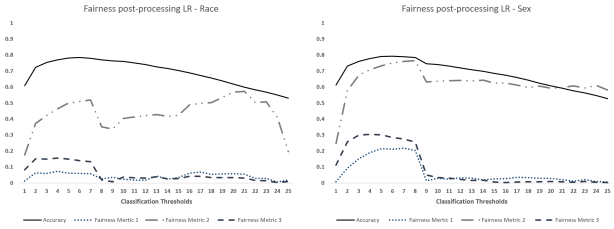


Fig. 3 Accuracy and fairness metrics trade-off for post-processing fairness enhancing LR on Race and Sex sensitive attributes

curacy and fairness constraints for different users. There is a need for a mechanism that can prioritize one constraint over the other based on the user requirements.

In our proposed mechanism, all the metrics (A, F1, F2, F3) act as agents and negotiate with each other based on the ratings they get from other agents and their trust sensitivities. All these agents have a goal. The accuracy agent has the goal of A = 1, and the fairness agents have the goal of F1 = 0, F2 = 0, and F3 = 0. In our scenarios, the distance is one-dimensional, but the distance metric can be used for multi-dimensional scenarios, as shown in Equation 3. Each agent aims to minimize their distance from the goal. For each solution, agents measure their distance from their goal. They rank the solutions from best to worst based on the distance to the goal. The best solution for an agent will be the solution with the minimum distance to the goal, and the worst solution will be the solution with the maximum distance from the goal. Each agent starts with their best solution and rates other agents' solutions based on their distance.

In the following rounds, each agent could (i) stay at the same solution or (ii) move to a new solution by increasing its distance from its goal and decreasing other agents' distance from its goals. And how much the agent sacrifices its own distance from its goal and increases other agents' distance to their goals depends on the rating it gets from other agents and its trust sensitivity. All the agents propose and rate each other solutions for multiple rounds.

To test our framework, we simulated eight rounds. In our experiments, we have considered trust sensitivity only for simplification purposes. If the trust sensitivity is

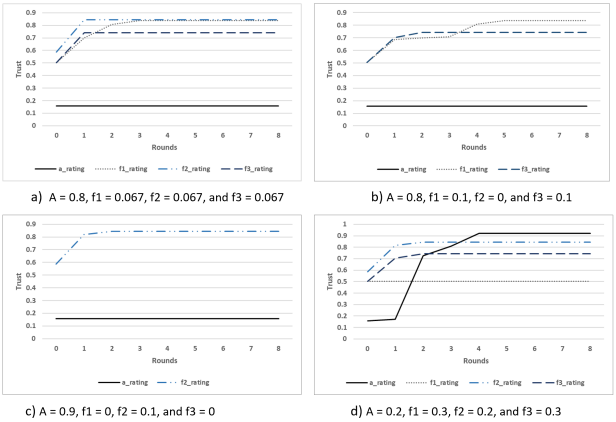


Fig. 4 Negotiation between different agents based on the weights and associated trust sensitivities for race attribute using post-processing LR algorithm

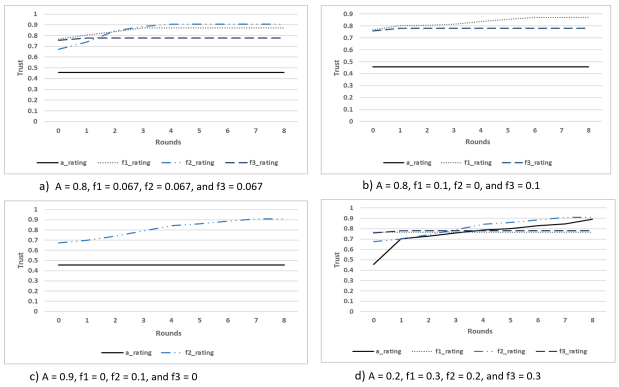


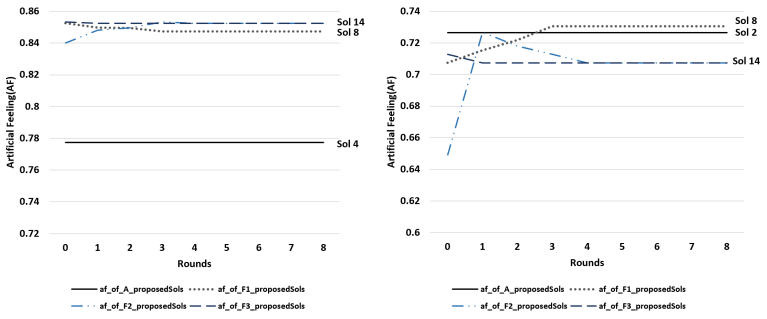
Fig. 5 Negotiation between different agents based on the weights and associated trust sensitivities for race attribute using post-processing SVM algorithm.

below 0.7, the agent does not move. If the trust sensitivity is between (0.7 - 0.9), the agent moves to its next best solution leading to an increase in the rating, and if the trust sensitivity is between (0.9 - 1.0), which means the agent is highly sensitive, the agent moves to the next second-best solution to increase its ratings quickly. We simulated different trust sensitivities, which are associated with the weights of the metrics. Figures 4 and 5 show the movement of the agents across rounds based on the weights assigned to them and their associated trust sensitivities for the race attribute. Figure 4 shows the graphs for the post-processing logistic regression algorithm, and Figure 5 shows the charts for the post-processing support vector machine algorithm. Each figure has four graphs for different metric weights. As seen in the figures, the movement of the accuracy agent in graph a) and graph d) is completely different.

In graph a), the accuracy agent is not sensitive given its higher weight hence not moving from its best solution. However, in graph d), the accuracy agent is negotiating and moving away from its best solution given its less weight. This shows how agents with different weights and sensitivities behave differently and lead to other solutions. Table 1 summarizes different solutions/classification thresholds reached by each agent after eight negotiation rounds for different metric weights for race and sex attributes.

Table 1 Solution reached for each agent after eight rounds of negotiation based on user-defined weights for post-processing fairness SVM and post-processing LR

Metrics Weights	Race Post-Processing SVM	Race Post-Processing LR	Sex Post-Processing SVM	Sex Post-Processing LR
a = 0.8, f1= 0.0667, f2= 0.0667, f3= 0.0667	*4,8,14,14*	6,9,8,8	5,8,13,2	6,10,9,9
a = 0.8, f1= 0.1, f2= 0.0, f3= 0.1	4,8,14	6,9,8	5,9,2	6,10,9
a = 0.9, f1= 0.0, f2= 0.1, f3= 0.0	4,14	6,8	5,13	6,9
a = 0.2, f1= 0.3, f2= 0.2, f3= 0.3	*2,8,14,14*	25,8,9,9	2,9,13,2	14,10,9,9



a) A = 0.8, f1 = 0.067, f2 = 0.067, and f3 = 0.067 b) A = 0.2, f1 = 0.3, f2 = 0.2, and f3 = 0.3

Fig. 6 Artificial Feeling (AF) during the rounds of negotiation for the proposed solutions of different agents.

Over the rounds of negotiation, we have calculated the ”artificial feeling” (AF) based on the user weights for each selected solution. Figure 6 illustrates how the artificial feeling(AF) changes for each agent in the negotiation rounds based on the solution proposal and user-defined weights. At the end of the negotiation, the solution with the highest Artificial Feeling (AF) is selected. As seen in Figure 6 graph a), Solution No. 14 is chosen for the given weight distribution because the AF of solution number 14 proposed by agents F2 and F3 is the highest, and based on graph b), Solution No. 8 is selected, which has the maximum AF value and was proposed by agent F1. This difference in solution numbers shows that the AF can capture the

context of the AI system based on the user's defined weights. We have also compared the AF for the same weight distribution but different algorithms. For a weight distribution of (A = 0.8, f1= 0.067, f2 = 0.067, f3 = 0.067), the maximum AF using SVM is 0.852, whereas the maximum AF using LR is 0.78. This shows that SVM performed better than LR for the same user requirement. So, our experiments show that the AF can capture different user contexts and be used as selection criteria for selecting the appropriate solution and algorithm for a given user context.

5 Conclusion

We have presented *"Artificial Conscience - Control Module"* framework to control the working of the AI system based on the user requirements. Based on the user-assigned weights, AI metrics-agents negotiate with each other using our trust engine and solves with maximum "Artificial Feeling." This type of framework can be applied to any AI system where multiple evaluation metrics are involved and when different users have different requirements from the AI system. Our framework uses the "Artificial Feeling" concept to select the best solution and algorithm for a given user requirement.

References

1. Baars, B.J.: In the theatre of consciousness. global workspace theory, a rigorous scientific theory of consciousness. Journal of consciousness Studies **4**(4), 292–309 (1997)
2. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al.: Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development **63**(4/5), 4–1 (2019)
3. Blum, L., Blum, M.: A theory of consciousness from a theoretical computer science perspective: Insights from the conscious turing machine. Proceedings of the National Academy of Sciences **119**(21), e2115934,119 (2022)
4. Chella, A., Manzotti, R.: Artificial consciousness. Andrews UK Limited (2013)
5. Dua, D., Graff, C.: Uci machine learning repository. university of california, school of information and computer science, irvine, ca (2019) (2019)
6. EC: Ethics guidelines for trustworthy ai (2018). URL https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
7. Information Technology– Artificial Intelligence – Overview of trustworthiness in artificial intelligence . Standard, International Organization for Standardization (2020)
8. Kaur, D., Uslu, S., Durresi, A.: Trust-based security mechanism for detecting clusters of fake users in social networks. In: Workshops of the International Conference on Advanced Information Networking and Applications, pp. 641–650. Springer (2019)
9. Kaur, D., Uslu, S., Durresi, A.: Requirements for trustworthy artificial intelligence–a review. In: International Conference on Network-Based Information Systems, pp. 105–115. Springer (2020)
10. Kaur, D., Uslu, S., Durresi, A.: Trustworthy ai explanations as an interface in medical diagnostic systems. In: International Conference on Network-Based Information Systems, pp. 119–130. Springer (2022)

11. Kaur, D., Uslu, S., Durresi, A., Badve, S., Dundar, M.: Trustworthy explainability acceptance: A new metric to measure the trustworthiness of interpretable ai medical diagnostic systems. In: Conference on Complex, Intelligent, and Software Intensive Systems, pp. 35–46. Springer (2021)
12. Kaur, D., Uslu, S., Durresi, A., Mohler, G., Carter, J.G.: Trust-based human-machine collaboration mechanism for predicting crimes. In: International Conference on Advanced Information Networking and Applications, pp. 603–616. Springer (2020)
13. Kaur, D., Uslu, S., Durresi, M., Durresi, A.: A geo-location and trust-based framework with community detection algorithms to filter attackers in 5g social networks. Wireless Networks pp. 1–9 (2022)
14. Kaur, D., Uslu, S., Rittichier, K.J., Durresi, A.: Trustworthy artificial intelligence: a review. ACM Computing Surveys (CSUR) **55**(2), 1–38 (2022)
15. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. Advances in neural information processing systems **30** (2017)
16. Ruan, Y., Zhang, P., Alfantoukh, L., Durresi, A.: Measurement theory-based trust management framework for online social communities. ACM Transactions on Internet Technology (TOIT) **17**(2), 1–24 (2017)
17. Russell, S.: Human Compatible: Artificial Intelligence and the Problem of Control. VIKING (2019)
18. Russell, S.J.: Rationality and intelligence. Artificial intelligence **94**(1-2), 57–77 (1997)
19. Russell, S.J., Subramanian, D.: Provably bounded-optimal agents. Journal of Artificial Intelligence Research **2**, 575–609 (1994)
20. Solms, M.: The Hidden Spring: A Journey to the Source of Consciousness. W. W. Norton & Company (2019)
21. Solms, M., Friston, K.: How and why consciousness arises: some considerations from physics and physiology. Journal of Consciousness Studies **25**(5-6), 202–238 (2018)
22. Sutcliffe, A.G., Wang, D., Dunbar, R.I.: Modelling the role of trust in social relationships. ACM Transactions on Internet Technology (TOIT) **15**(4), 1–24 (2015)
23. Uslu, S., Kaur, D., Rivera, S.J., Durresi, A., Babbar-Sebens, M.: Decision support system using trust planning among food-energy-water actors. In: International Conference on Advanced Information Networking and Applications, pp. 1169–1180. Springer (2019)
24. Uslu, S., Kaur, D., Rivera, S.J., Durresi, A., Babbar-Sebens, M.: Trust-based game-theoretical decision making for food-energy-water management. In: International Conference on Broadband and Wireless Computing, Communication and Applications, pp. 125–136. Springer (2019)
25. Uslu, S., Kaur, D., Rivera, S.J., Durresi, A., Babbar-Sebens, M.: Trust-based decision making for food-energy-water actors. In: International Conference on Advanced Information Networking and Applications, pp. 591–602. Springer (2020)
26. Uslu, S., Kaur, D., Rivera, S.J., Durresi, A., Babbar-Sebens, M., Tilt, J.H.: Control theoretical modeling of trust-based decision making in food-energy-water management. In: Conference on Complex, Intelligent, and Software Intensive Systems, pp. 97–107. Springer (2020)
27. Uslu, S., Kaur, D., Rivera, S.J., Durresi, A., Babbar-Sebens, M., Tilt, J.H.: A trustworthy human–machine framework for collective decision making in food–energy–water management: The role of trust sensitivity. Knowledge-Based Systems **213**, 106,683 (2021)
28. Uslu, S., Kaur, D., Rivera, S.J., Durresi, A., Durresi, M., Babbar-Sebens, M.: Trustworthy acceptance: A new metric for trustworthy artificial intelligence used in decision making in food–energy–water sectors. In: International Conference on Advanced Information Networking and Applications, pp. 208–219. Springer (2021)