Towards Trusting the Ethical Evolution of Autonomous Dynamic Ecosystems

Emilia Cioroaica
Safety Engineering
Fraunhofer IESE
Kaiserslautern, Germany
emilia.cioroaica@iese.fraunhofer.de

Barbora Buhnova
Faculty of Informatics
Masaryk University
Brno, Czech Republic
buhnova@mail.muni.cz

Emrah Tomur
Ericsson Research
Ericsson
Istanbul, Turkey
emrah.tomur@ericsson.com

***Abstract*—Until recently, systems and networks have been designed to implement established actions within known contexts. However, gaining the human trust in system behavior requires development of artificial ethical agents proactively acting outside fixed context boundaries for mitigating dangerous situations in which other interacting entities find themselves. A proactive altruistic behavior oriented towards removing danger needs to rely on predictive awareness of a dangerous situation.**

Different that current approaches for designing cognitive architectures, in this paper, we introduce a method that enables the creation of artificial altruistic trusted behavior together with an architecture of the framework that enables its implementation.

***Index Terms*—Trust, Digital Ecosystems, Simulation, Ethics, Morality, Artificial Agent, AI**

I. Introduction

With the capability of modelling structures of various entities by learning from large data sets, AI (Artificial Intelligence) systems are becoming the enablers of intelligent networks, where allocation and distribution of workload are performed based on computation costs and device constraints. By integrating self-governed subsystems of networked functions within various application domains, a new type of digital ecosystem is emerging. Autonomous dynamic ecosystems have an intelligence distributed across multiple digital agents that decide on task allocation and resource consumption in a decentralized manner. In a scenario where autonomous intelligent networks and interconnected systems enable blind people to navigate in open context, AI agents integrated within networks need to ensure the necessary computation power for processing large quantities of data at high rates while at the same time, they are required to account for sustainable energy usage. To meet such requirements, 6G technical developments is exploring upper mmWave spectrums (100-300Ghz) and deployment of new distributed technology such as re-configurable Intelligent Surface (RIS) and ML (Machine Learning) aided network operation that can harness highly complex calculations and environmental modelling [1].

In the above scenario, assuring a safe and ethical operation of AI-enabled networks becomes a necessity for gaining the societal trust. Emerging evidence that pinpoints the negative impact of technological solutions on humans and society [2] elevates the requirements of digital moral evolution. Technology is capable of providing development for virtually endless applications, however, it is often overlooked or neglected that these systems can be exploited in unethical directions [3]. Even if scholars are becoming aware of the double-edged sword of technological progress [4], the current practices only aim at protecting humans by considering malicious attacks on systems, without considering attacks directed towards exploitation of human behavior inherent to these systems.

By incorporating AI within interconnected networks and system processes, we implicitly delegate moral responsibilities to these systems and to the emerging ecosystems. In our opinion, a trusted evolution of emerging autonomous dynamic ecosystems can only be possible through integration of mechanisms capable of foreseeing the effects of actions as a basis for trusted ethical reactions. In the past, we have explored the mechanism of predictive simulation in the context of technical trust alone in [5]. In this paper, we are elevating the trust consideration through integration of ethical properties of systems and ecosystems for which we are uplifting the mechanism of predictive simulation for enabling ethical evaluation of artificial altruistic behavior. By enabling the creation of predictive awareness, the predictive simulation supports the process of ensuring moral operation of AI-controlled systems within autonomous dynamic ecosystems.

In what follows, Section II presents a summary of emergent development trends together with a holistic definition of trust that accounts of the variety of entities within autonomous dynamic ecosystems. Section III introduces our method for building trust in the artificial ethical behavior of interconnected systems and networks. Section IV presents the concept of our platform used to dynamically evaluate the ethical aspects of an artificial action and Section V presents the conclusions together with the future research roadmap.

II. Emerging Trends

A. Formation of Internet of Senses

The interconnection within digital world, initially a characteristic of system components, has transitioned towards SoS (System of Systems) and recently has been uplifted to be characteristic of dynamic interconnections between various actors such as businesses, developers, systems and system components within *digital ecosystems* [6]. Emerging technological development is raising the inter-connectivity to yet another level. Supported by 6G technology, the interconnection between the digital world characterized by information, algorithms and organisms emerges towards formation of the *Internet of Senses* [7].

It is envisioned that this new ecosystem will connect all types of intelligence, including crowd intelligence and artificial intelligence together with existing networks of networks, overall leading to a convergence of data, connectivity and specialized platforms used for creating new businesses, new stakeholders and new regulations.

B. Towards a holistic trust-evaluation process

Given the scope of connecting multiple entities with different characteristics such as humans, physical systems, digital assets, businesses, and networks acting in different domains, emergent solutions for building trust need to employ a holistic understanding of the term in encompassing domains of automation, networking, economics and human relationships.

In the literature, the term of *trust* transcends in its understanding between different domain. In social science *trust* is related to actions that account for uncertainty and ignorance [8], becomes influenced by personal and moral relationship between two entities [9] in philosophy, and is linked to achieving a calculated incentive [10] in economics. Overall, in the domains of sociology, philosophy and economics, *trust* is regarded as a quantified subjective attitude of an action which under psychological considerations becomes the outcome of a process acquired from experiences [11].

Consequently, for the development of sustainable solutions within autonomous digital ecosystem, the understanding of trust traditionally associated with reliable behavior of a system [12] needs to elevate based on prerequisites of automated processes that consider an agent's action of achieving the goal of another agent [13] based on the belief of trustworthiness derived from the willingness to reciprocate a cooperation [14].

While humans have a native perception of the concept of trust, machines need to be enhanced with the possibility of reasoning about it. Consequently, interconnected systems and networks need to employ mechanism that enable gathering of evidence to support an internal claim of trust or/and distrust preferably in a way that enables the system to react internally through reconfiguration and externally through proactive reaction to foreseen negative consequences of an action of itself and/or of interconnected entities.

C. The need and challenges of Ethical supervision

While there is still an ongoing debate over the capacity of machines to be ethical, in the field of Machine Ethics, philosophers and engineers are starting to unite their forces for developing such machines [15], [16]. In the traditional engineering process, ethical aspects are considered as an extension of the safety envelope derived from a design time-safety engineering process, which accounts of ethical decisions along the system design process leading to the creation of implicit ethical agents. However, emergent evidence shows that the operation of AI-driven machines can unpredictably evolve and require runtime supervision as well [17].

Transitioning from the traditional domain of robotics where integration of AI solutions has lead to the consideration of ”minimally ethical robot” [18] capable of implementing actions whose consequences are significant in moral deliberations, AI-driven networks will soon drive the need for ethical considerations and will consequently require implementation of dedicated ethical measures. It is well known that in judging a machine's morality, ethical decisions that are difficult for humans are equally challenging for machines and even more challenging for the designers of these machines who need to account of human intolerance to a system error which is greater than the intolerance to errors performed by other humans.

III. Method

By leveraging the ethical considerations of AI-powered robots [19] to AI-powered networks of networks within the Internet-of-Senses, we consider the emerging network systems being *ethical impact agents* as well. Between the multitude of attributions, AI agents incorporated within existing networks are regarded as black boxes that manifest critical low-level decision on limited resource usage. The reasoning behind such decisions are typically not known and for gaining the trust within an ecosystem, these decisions need to be evaluated on various levels.

Further on, for enabling the development of an ethical altruistic behavior we refer to the characteristics of human ethical behavior. Aside from showing an altruistic behavior, humans are capable of consciously reflecting on actions and justify the morality of these actions. Consequently, an autonomous digital ecosystem and encompassing AI-powered ecosystem components (e.g. systems, networks) need to account of its own actions as well as of actions of interconnected digital assets and systems that have an ethical impact. We consider the degree of a system's capability to be an ethical agent in accordance with the level of development that leads to ethical outcomes characterized by an artificial altruistic behavior reflected in the implementation of an action that prevents a hazardous situation outside the scope of the current operation. For example, within an autonomous dynamic ecosystem comprising of a multitude of interconnected systems and networks of networks, an AI-powered smart home connected to a smart grid can foresee a dangerous situation of a blind person and can trigger events for preventing a hazard.

For enabling a trusted execution of events in the above scenario, we elevate the mechanism of predictive simulation described in [5] in the domain of intelligent autonomous ecosystems formed by heterogeneous complex entities (systems, network, humans). Assuming an AI to be a Machine Learning Algorithms (ML) used for learning ethical actions within an ecosystem, we see the necessity of applying the principle of predictive leverage [20] on context change. This means that the explanatory principle used by the ML algorithm shall be tested in the new context, when the problem or the parameters are changed. Because the context changes dynamically at runtime, we suggest that the testing needs to be carried out at runtime as well and requires deployment of platforms capable of performing a dynamic runtime evaluation of emergent behavior.

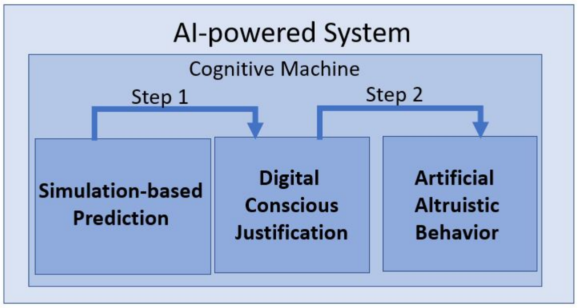
AI-powered System
Cognitive Machine
Step 1
Step 2
Simulation-based Prediction
Digital Conscious Justification
Artificial Altruistic Behavior

Fig. 1. The Method

In our vision depicted in Fig. 1, an AI-powered system, which can be a network and/or a communicating cyber-physical system contains a *Cognitive Machine* that drives the *Artificial Altruistic Behavior* in two steps. In *Step 1*, the Cognitive Machine performs a *Simulation-based prediction* by dynamically executing abstract models of the system and interacting entities within an established collaboration context. A this moment, the operational context of a system is enhanced with context information of the collaborating entities. The predictive simulation enables the implementation of the predictive leverage principle by enabling a predictive evaluation of an action within a dynamically created context. In this way, when the problem or the parameters of the intended action are changed, it will be possible to virtually test if the ethical principle still holds and it is not endangering an interacting entity or the environment. Most networks and systems are implicit agents, in the sense that they are designed according to ethical aspects implemented alongside their safety engineering process. The degree of explicit ethical agents is defined by the level on which those systems have learned or defined ethical rules and therefore require a cognitive machinery that considers these rules in the acting decision in a given context. The concept of predictive simulation employed within a cognitive machinery enables an evidence-based justification of the proactive artificial act of intervention to prevent harm within an overall process of ethical reasoning.

Within a simulation-based prediction, abstract models of the system itself and of the interacting entities are executed for providing a dynamic evidence of trust based on which a *Digital Conscious Justification* is formed. As we will exemplify in Section IV, the Digital Consciousness will perform dynamic evaluation of goals, driven by artificial decisions, which are influenced by ethical principles.

In the *Step 2*, The Digital Conscious Justification leads to the expression of an *Artificial Altruistic Behavior* which is a consequential ethic behavior of an entity who carries extra actions for avoiding hazardous situations of another entity (human, system or network). This behavior is based on the capability to anticipate consequences of the other entity's actions as well as capability and will of acting in order to prevent a possible calamity. In this phase, the process of attributing moral responsibilities to an artificial entity is based on engineering and technical considerations that ensure the entity's capability to anticipate and react in a trustworthy manner. The mechanism of predictive simulation supports a system's capacity to anticipate reactions within an internal probating period during which ethical decisions are evaluated and reviewed before being exposed in the real world.

The method we are proposing encompasses multiple challenges such as: a) considerations of rules to learn, b) choice of training cases that can be scrutinized by a panel of multiple and various stakeholders, c) transparency of the ethical decision-making process which can require logging of selected information derived from data collected at the operational level, d) creation of dynamic models that reflect the new operational context, e) dynamic runtime formation of abstract data-driven abstract models that characterize a new environment, f) deployment of a mechanism that enables the dynamic trust evaluation of ethical outcomes. In the next section we introduce our designed solution to the challenge f), as a step towards enabling the applicability of our approach.

IV. Platform

Within an autonomous digital ecosystem, an entity decides without human intervention how to respond to inputs. In ”supervised autonomy” [21], autonomous systems are subject to human monitoring or supervision. However the low-level decisions are made by the system itself. Therefore, the evaluation and prediction needs to be performed at the operational level as well. Artificial entities that act autonomously and according to ethical principles take decisions either according to predetermined principles or learn these principles from observed decisions. When ethical decisions are learned from data consisting of a set of moral judgements, a general principle of ethics is derived for the corresponding artificial entity. As stated before, this principle shall have predictive leverage in the sense that it shall be tested as soon as the problem/parameters change.

Current approaches that judge the morality of a machine use the human behavior as a threshold [22]. However, it is well-known that humans are more sensitive to errors made by machines than to errors made by humans. Therefore, such a threshold can be too low for deriving the behavior of a machine that can be ethically trusted in a society. Instead, in our work we propose an elevation of the current concepts designed to avoid unethical system outcomes based on reasoning about obligations and permissions [15] by incorporating a runtime-based prediction within a holistic reasoning process of trust.

For this, in Fig 2 we introduce the concept of a platform that enables the dynamic testing of an ethical principle within an autonomous dynamic ecosystem. An *Autonomous Dynamic Ecosystem* consists of multiple *Actor*s such as *User*s, *Businesses*, *Developer*s within organizations and *Artificial Ethical Impact Agent*s, which are artificial entities such as *System*s, *Network*s and other *Digital Asset*s which can be software components and software smart agents. For assuring an ethical

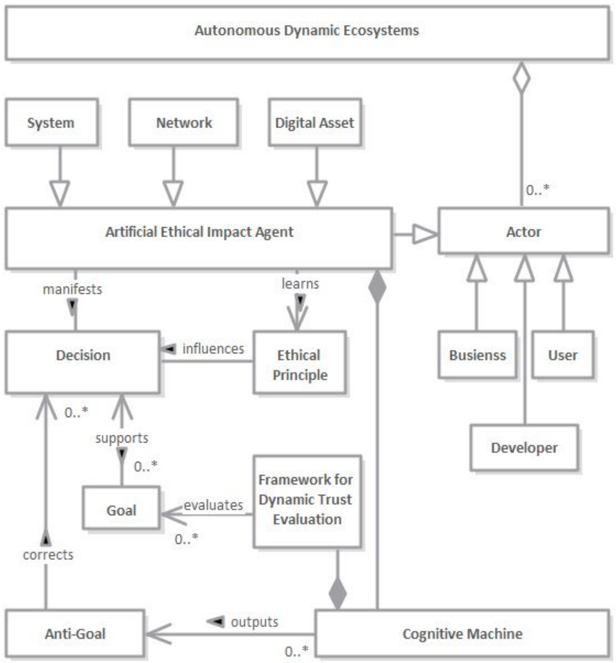

Fig. 2. Metamodel of the Platform

behavior, the predictive runtime evaluation needs to account of a) the *Decision*s manifested by the *Ethical Impact Agent*s and b) the *Ethical Principle* that is learned. A digital *Cognitive Machine* that contains a *Framework for Dynamic Trust Evaluation* dynamically evaluates the *Goal*s supported by the *Decision*s and outputs detected deviations in terms of *Anti-Goal*s that will correct a Decision in a virtual environment, before it takes effects in the real world.

V. Conclusions and future work

Human expectations of a system's trustworthy behavior are already higher when compared to human expectations of trusted behavior of another human. Technological advances and the vision to develop explicit ethical agents raise the human expectation on ethical machines even higher, making their ethical governance even more challenging. In this paper we've introduced a vision for building ethical trust by elevating the concept of predictive simulation for enabling evaluation of ethical outcome decisions in a predictive virtual environment, as a prerequisite for enabling a prompt ethical reaction. Different than current approaches for designing minimally ethical agents that focus on construction of cognitive architectures that support the core processes of making ethical choices, the predictive simulation, part of a cognitive machinery enables a proactive reasoning and ethical trustworthy to possible harmful situations.

In addition to the research directions mentioned at the end of Section III, for enabling the further development of an ecosystem's awareness to harmful consequences of internal actions, multiple engineering challenges must be solved, such as: a) development of capabilities to recognize situations in which actions have ethical impact, b) development of capabilities to perceive the nature of risk when interacting with humans, c) development of capability to perceive risk in a new context, and d) development of capability to adapt the risk computation in dynamic changes of context.

Acknowledgment

This work was supported by the project BIECO (www.bieco.org) that received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 952702, and from ERDF/ESF "CyberSecurity, CyberCrime and Critical Information Infrastructures Center of Excellence" (No. CZ.02.1.01/0.0/0.0/16_019/0000822).

References

[1] "https://hexa-x.eu/wp-content/uploads/2021/09/Hexa-X-D4.1$_{v}1.0.pdf,''$ 2021.

[2] T. Zhang, "Three essays on the economics of cybersecurity," Ph.D. dissertation, 2020.

[3] S. Aral and D. Eckles, "Protecting elections from social media manipulation," *Science*, vol. 365, no. 6456, pp. 858–861, 2019.

[4] R. Rathi, "Effect of cambridge analytica's facebook ads on the 2016 us presidential election," *Towards Data Science*, 2019.

[5] E. Cioroaica, T. Kuhn, and B. Buhnova, "(Do not) trust in ecosystems," in *Proceedings of the 41st International Conference on Software Engineering: New Ideas and Emerging Results*. IEEE, 2019, pp. 9–12.

[6] E. Cioroaica, A. Purohit, B. Buhnova, and D. Schneider, "Goals within trust-based digital ecosystems," in *2021 IEEE/ACM (SESoS/WDES)*. IEEE, 2021, pp. 1–7.

[7] "Hexa Deliverable D1.2 Expanded 6G vision, use cases and societal values."

[8] D. Gambetta *et al.*, "Can we trust trust," *Trust: Making and breaking cooperative relations*, vol. 13, pp. 213–237, 2000.

[9] O. Lagerspetz, *Trust: The tacit demand*. Springer Science & Business Media, 1998, vol. 1.

[10] H. S. James Jr, "The trust paradox: a survey of economic inquiries into the nature of trust and trustworthiness," *Journal of Economic Behavior & Organization*, vol. 47, no. 3, pp. 291–307, 2002.

[11] J. B. Rotter, "Interpersonal trust, trustworthiness, and gullibility." *American psychologist*, vol. 35, no. 1, p. 1, 1980.

[12] J.-H. Cho, A. Swami, and I.-R. Chen, "A survey on trust management for mobile ad hoc networks," *IEEE Communications Surveys Tutorials*, vol. 13, no. 4, pp. 562–583, 2011.

[13] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[14] B. L. Slantchev, "Trust and mistrust in international relations," *Perspectives on Politics*, vol. 4, no. 3, p. 632–633, 2006.

[15] S. Bringsjord, N. Sundar G., D. Thero, and M. Si, "Akratic robots and the computational logic thereof," in *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering*, 2014, pp. 1–8.

[16] G. M. Briggs and M. Scheutz, ""sorry, i can't do that": Developing mechanisms to appropriately reject directives in human-robot interactions," in *2015 AAAI fall symposium series*, 2015.

[17] M. Lewis, D. Yarats, Y. N. Dauphin, D. Parikh, and D. Batra, "Deal or no deal? end-to-end learning for negotiation dialogues," *arXiv preprint arXiv:1706.05125*, 2017.

[18] A. F. Winfield, C. Blum, and W. Liu, "Towards an ethical robot: internal models, consequences and ethical action selection," in *Conference towards autonomous robotic systems*. Springer, 2014, pp. 85–96.

[19] J. Moor, "The nature, importance, and difficulty of machine ethics," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 18–21, 2006.

[20] A. F. Winfield, K. Michael, J. Pitt, and V. Evers, "Machine ethics: The design and governance of ethical ai and autonomous systems [scanning the issue]," *Proceedings of the IEEE*, vol. 107, no. 3, pp. 509–517, 2019.

[21] P. Sexton, L. S. Levy, K. S. Willeford, M. G. Barnum, G. Gardner, M. S. Guyer, and A. L. Fincher, "Supervised autonomy," *Athletic Training Education Journal*, vol. 4, no. 1, pp. 14–18, 2009.

[22] C. Allen, G. Varner, and J. Zinser, “Prolegomena to any future artificial moral agent,” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 12, no. 3, pp. 251–261, 2000.