The Concept of Ethical Digital Identities

Emilia Cioroaica*, Barbora Buhnova†, Frank Jacobi‡ and Daniel Schneider*
*Fraunhofer IESE, Kaiserslautern, Germany
†Masaryk University, Brno, Czech Republic
‡formitas AG, Aachen, Germany
*{emilia.cioroaica, daniel.schneider}@iese.fraunhofer.de
†{buhnova@mail.muni.cz} ‡{fj@formitas.de}

***Abstract*—Dynamic changes within the cyberspace are greatly impacting human lives and our societies. Emerging evidence indicates that without an ethical overlook on technological progress, intelligent solutions created to improve and enhance our lives can easily be turned against humankind. In complex AI-socio-technical ecosystems where humans, AI (Artificial Intelligence) and systems interact without a common language for building trust, this paper introduces a methodological concept of Ethical Digital Identities for supporting the ethical evaluation of intelligent digital assets.**

***Index Terms*—Ethics, Morality, Engineering, AI, Trust, Process, Architecture**

I. Introduction

Emerging moral vulnerabilities of AI-powered systems, and autonomous AI-controlled systems in particular, become increasingly harder to fix as new evidence shows that battlefields are transitioning from the cyberspace to the human mind [1]. Technology provides development for virtually endless applications, however, it is often overlooked or neglected that through psychological, social and neurological implications, ethical vulnerabilities can be exploited via interference on human behavior [2]. Even if scholars are becoming aware of the double edged sword of technological progress [3], the current practices only aim to protect humans by considering malicious attacks on systems, without considering attacks that exploit human psychology that are inherent to these systems.

Currently, the effort of understanding the human psychology is directed towards boosting business growth. Consequently, AI has transitioned into an instrument that negatively impacts human behavior [4]. If this trend continues, we will soon face powerful mistrust, AI misuse and/or an extensive disuse of digital solutions which may be regarded as ”evil”.

Transitioning from the domain of information systems, which until recently was the only one capable of providing enough resources for complex and ultra fast computations required by AI, the emerging technological development in embedded systems is paving the way towards intelligent control within the domains of automotive and healthcare. In the automotive domain, in particular, an increasing number of vehicle components are outfitted with AI to enhance their functionality. For example, infotainment systems are enhanced to suite the driver [5], navigation systems select routes based on the preferences of passengers and climate control systems create the most suitable environment for passengers [6]. And unfortunately, these safety-critical domains can only partially learn lessons from the process of engineering information systems which mainly deal with data in its aggregated form.

Regarding data handling, information systems have provided evidence that when a society loses its attention on what kind of information it circulates, it can be endangered by a lot of potential threats, at the organizational or even at the national level [7]. Consequently, the domain of safety-critical systems currently hosts methods such as explainable AI together and methods for ensuring that the data sets from which a system learns its ethical behavior cannot be interfered with.

However, systems from safety-critical domains pose a major distinction through their capability of physically actuating and changing their immediate environment. The current research landscape of machine ethics of autonomous systems classifies the technical implementation of moral concerns into a range of *ethical agents* [8], which vary on the degree of ethical reasoning, with the highest level of ethical agent witnessed today being the *explicit ethical agents* that never-the-less has only very small ethical considerations [9].

For developing the next level of ethical agents which would be *fully ethical agents* [8] capable to make explicit moral judgements and justify them, in this paper we introduce the concept of EDI (Ethical Digital Identities) as methodological transition that involves the contribution of multiple stakeholders in making and justifying the moral judgements, achieving in this way a societal-driven justification of morality.

In what follows, Section II presents social concerns that converge towards the need of taking immediate actions in the wide field of emerging societal and technological evolution presented in Section III. Section IV opens up the discussion on ethical assurance of digital assets through the introduction of a structured process centered on the notion of EDIs. Section V presents conclusive discussion and a future research roadmap.

II. Emerging Threats

A. Battle Threats

In human history, land territories were the initial battlefield. As damages became obvious, moral guidelines were put in place to stop it. Then the digital developments have uplifted the quest into the cyberspace [10] and organizations around the world started to join their efforts and invest tremendous monetary resources in fixing the threats on cybersecurity [1]. Yet, there is an emerging field of battle that gets little attention in the engineering society: the human mind. While social sciences already regard human behavior to be a chain of factors that lead to predictable actions [11], only business attention is currently directed towards exploring this fact [4]. No engineering solution for fixing the problem exists at the moment. Social regulations are subject to governmental entities which due to fast rotations cannot commit to long-term engineering solutions. In the worst case, governmental entities succeed each other with contradicting views and priorities on socio-technological development. And while social sciences can pinpoint to the problem quite well [3], they lack the instruments for designing large scale solutions. This is due to the subjective and complex human nature which cannot be safeguarded against its own vulnerabilities towards manipulation. As a matter of fact, any evidence in this sense is disregarded by the victim in what is known as conformation bias [12]. Equipped with instruments that enable separation of concerns, we believe that engineers together with social scientists can design and construct solutions to the ethical problems of AI-human coexisting societies.

B. *Risks and attacks*

It is long known that personality types and individual cognition greatly influence the human risk perception [13]. This further on influences the way in which information is assimilated and dictates the ultimate human behavior [14]. In the hands of a malicious attacker, stimulation for a risk taking behavior can easily become an instrument for safety attacks. For example, in the automotive domain, an AI component which is controlling the air suspension system for the wheels could induce strange lifting patterns that imitate a failure or a rocky road. A human who is the ultimate control entity of the vehicle can then undertake a risky behavior, fueled by the unconscious need to maintain physiological arousal while focusing on the rewards [13]. With psychological instruments at hands, such as supraliminal and subliminal priming [15], it is only a matter of time, until intended risky behavior can be induced to humans through creation of false memories or sensations [16]. Besides immediate safety implications, such attacks can have long lasting effects. The psychological and ethological mechanism called *(Limbic) Imprinting* [17] shows how highly emotional experiences can create deep neurological connections, influence behavior and physical health even for generations.

III. Emerging Transitions

A. *Societal transitions*

Adoption of AI solutions is envisioned to uplift the human responsibilities in the emergent *feeling economy* [18]. The repetitive and analytical human cognitive processes are increasingly assigned to AI components, leaving human workers to address more interpersonal, empathetic and ethical tasks.

To support the social transition in the extreme labour displacement envisioned to be caused by the AI integration into our societies [19], the nature of jobs people are asked to perform needs to change. When automatic and repetitive tasks will be the core responsibility of automated processes, humans need to be given the possibility to exercise their human attributes for the benefit of societal growth. For example, when intelligent robos will do the jobs of the humans, humans will be free to think and develop social connections or engage in the moral development of intelligent systems. However, a complete and instant switch is almost impossible, as people tend to misuse or disuse technologies that they do not know [20].

A solution to this problem is to add humans in the evolution cycle of AI systems, eventually in return of monetary rewards. Besides the social and economic benefits of creating new jobs, when people are engaged into evolution processes of technological solutions, they tend to trust these solutions much more. And technological support for crowd source implementation such as the Mechanical Turk [21], that start to emerge on the open market, can as well support the societal-driven moral development of systems.

B. *Technological Transitions*

For a long time, technological assessment of engineering solutions has been performed according to implications on safety risks, morality being regarded as an ”extended safety envelope” [9]. The recent enforcement of ethical evolution of AI-based systems, provided by the European Commission through appointment of High-Level Expert Group on Artificial Intelligence in a series of ethical guidelines [22], are currently shaping the research landscape of systems’ design with more detailed requirements on responsible technological development [23]. Technical standards such as [24] and [25] are emerging and explicit normative encoding are raising the public awareness to these principles. The research results are translated in defining the needed steps of identifying moral concerns [26] and mapping them to the technological solutions through provision of principles [27] or checklists [28].

No engineering solution currently exists for dynamically safeguarding the ethical development of current and emerging new AI technologies. In this work we open up the discussion on possible process implementations by proposing a structured methodological approach that assigns the supervision of ethical progress to experts while being exercised by the general population.

IV. The Concept for Ethical Evaluation

In this section we introduce a conceptual method for ethical evaluation of digital assets centered around the concept of Ethical Digital Identities (EDI).

A. *Methodological Concept*

In Fig. 1 we introduce a structured view of the methodological concept for the envisioned Ethical Digital Identities aimed to support the leveraging of current and future human-technical processes of trust evaluation through integration of ethical considerations. For this, a *Digital asset* subject to be introduced in the cyberspace is evaluated against an *Ecological Principle*. The Ecological principle imposes an ecological perspective by considering common benefits of the asset provider, user and the environment in accordance to guidelines specified in [23]. The Ecological Principle is the central mechanism that converges businesses into safe guarding the ethical evolution of the assets. The digital asset can be an *AI solution* that enables process automation or a piece of *Information* introduced into the cyberspace.

From the business perspective, the ecological principle is evaluated in scenarios that describe the *Business Gains*. From the social perspective, the digital asset is evaluated against *Social Gain* which can be an *Individual Gain* and/or an *Environmental Gain*. Business and ethical experts define evaluations scenarios according to business and ethical perspectives such as the ones in [22]. Focusing on the social implications of digital assets, in this paper we exemplify the evaluation mechanisms from the social perspective only. Such evaluation scenarios can be defined by experts and exercised by the general population in return of monetary rewards.

The Social Gain is evaluated by multiple *Ethical Observer*s which form a *Virtual Commission* that provides certified *Accreditation*. Similar to the certification of safety-critical systems, such accreditation is envisioned to be in the responsibility of a group of authorized certified social experts. The suitability to ethical accreditation can follow a similar approach used for certifying safety-critical systems, but for moral considerations [28].

The accreditation entity generates the structure of an *Ethical Digital Identity* (EDI), similar to Digital Identities [29]. The EDI can be customized to the digital asset which is under ethical evaluation and it is continuously updated during the lifetime of its corresponding digital asset. The ethical identity is updated according to *Evidence* that supports the creation of a *Holistic Perspective* of its moral dimensions. The balanced perspective of information is assisted to integration of both *Supporting Evidence* and *Counter Evidence*. Supporting evidence can be generated from successful evaluation scenarios whereas counter evidence can be generated from cases in which a particular component cannot be trusted.

When humans take part in the evolution process of an AI component, information can efficiently be transmitted through mechanisms that help human brains to retain information. Such mechanisms are implemented by the *Efficient Visual Description* components. Human brain retains and judges information based on immediate visual effects and past experiences. Therefore, the display of ethical information needs to be performed according to suitable psychological mechanisms decided by social and user experience experts. Finally an EDI has a *Digital Signature*, which enables formation of ethically trusted communication channels in the cyberspace. Through digital signatures, providers of digital assets can, for example, safeguard the ethical quality of an AI component and receive a positive social reputation.

Through the EDI, a digital asset becomes a living structure. It receives an ethical identity issued by certified ethical and expert authorities. For example, when the digital asset is a piece of information planned to be displayed in AR (Augmented Reality) application within a vehicle, the *Virtual Commission* can consist of ethical observers, such as experts in the psychology, sociology and neurology. Wrong information displayed on AR devices within the car can lead to the fear that the car is acting strangely and force the actuation of a driver who is in a nervous state.

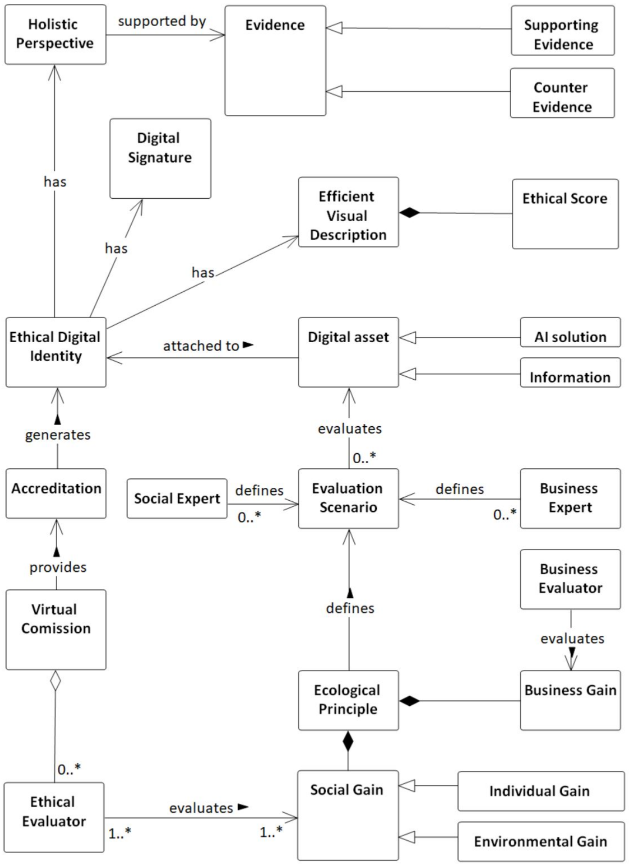

Fig. 1. The methodological concept for ethical process evaluation centered on EDIs

B. *Requirements for Ethical Digital Identities*

Ensuring morality of intelligent digital assets is the key to unlock their full potential in our societies, enable industries to develop confident business models and nurture their social uptake. For enabling this vision, our concept of Ethical Digital Identities (EDI) elevates the concept of Digital Identities [29] to the domain of ethics for safety-critical systems, and sets the basis for safe-guarding the ethical evolution of intelligent digital assets. At its core, a Digital Identity is defined as "*the data that uniquely describes a person or a thing and contains information about the subject's relationships*" [29]. Similarly, an EDI needs to be uniquely assigned to a digital asset and needs to transmit to the consumer and moral certification authorities evidence of its psychological implications. An EDI should bring the concerns of social experts into the domain of cyberspace while enabling dynamic exercise and evaluation of ethical and social concerns by non-experts with the possibility of engaging the large population in return of monetary incentives.

Through EDIs assets can become living identities with traceable evidence of moral implications evaluated from experts' social concerns. The Structure of an EDI should be designed by social-science experts, implemented by engineers, issued in the digital space together with its corresponding digital asset and continuously maintained during the complete lifetime of its corresponding digital asset. Design of EDIs should be integrated into current engineering practices and should also enable an ethical start of new technological developments.

V. Conclusions and Future Research Roadmap

Emerging evidence from domains where AI is largely used shows that psychological aspects are at least as important as technical aspects in enabling a sustainable digital evolution of our societies. The social concerns of technology providers are addressed in this paper through the introduction of a concept that enables the ethical evaluation of digital assets. Used for boosting the business growth, current technological use of solutions like Mechanical Turk [21], can also be transferred into supporting the morally supervised evolution of digital assets. The new concept of EDI (Ethical Digital Identity) has been introduced in the center of a structured methodological concept that enables the integration of its core principles into ongoing as well as into new processes. By pointing to processes and literature emerging in the domain of ethical technology, the EDIs support integration of well stated practices and ethical considerations.

Our approach for shaping a moral evolution of AI technology within safety-critical domains encompasses related research directions such as (a) design of protective mechanisms for safeguarding the Virtual Commission from being itself subject to malicious attacks, misuse, or manipulation-lead distrust, (b) exemplification of concrete scenarios for the use of EDIs in different contexts, (c) validation of concept's completeness in different contexts, (d) detailing of the process in concrete scenarios, and (e) design of an overall trust assurance case that builds on ethical evidence along side the technical evidence.

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

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

[3] S. Scherr and A. Brunet, "Differential influences of depression and personality traits on the use of facebook," *Social Media+ Society*, vol. 3, no. 1, p. 2056305117698495, 2017.

[4] O. Turel, "An empirical examination of the "vicious cycle" of facebook addiction," *Journal of Computer Information Systems*, vol. 55, no. 3, pp. 83–91, 2015.

[5] A. Gaffar and S. Monjezi, "Using artificial intelligence to automatically customize modern car infotainment systems," in *Proceedings on the International Conference on Artificial Intelligence (ICAI)*. The Steering Committee of The World Congress in Computer Science, Computer ..., 2016, p. 151.

[6] A. Rosenfeld, A. Azaria, S. Kraus, C. V. Goldman, and O. Tsimhoni, "Adaptive advice in automobile climate control systems," in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, 2015, pp. 543–551.

[7] D. Helbing, B. S. Frey, G. Gigerenzer, E. Hafen, M. Hagner, Y. Hofstetter, J. Van Den Hoven, R. V. Zicari, and A. Zwitter, "Will democracy survive big data and artificial intelligence?" in *Towards digital enlightenment*. Springer, 2019, pp. 73–98.

[8] J. H. Moor, "The nature, importance, and difficulty of machine ethics," *Machine ethics*, pp. 13–20, 2011.

[9] 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.

[10] P. Chithaluru, R. Tanwar, and S. Kumar, "Cyber-attacks and their impact on real life: What are real-life cyber-attacks, how do they affect real life and what should we do about them?" *Information Security and Optimization*, p. 61, 2020.

[11] M. Schrijvers, T. Janssen, O. Fialho, and G. Rijlaarsdam, "Gaining insight into human nature: A review of literature classroom intervention studies," *Review of Educational Research*, vol. 89, no. 1, pp. 3–45, 2019.

[12] Sude, Pearson, and Knobloch-Westerwick, "Self-expression just a click away: Source interactivity impacts on confirmation bias and political attitudes," *Computers in Human Behavior*, vol. 114, 2021.

[13] R. Lion and R. M. Meertens, "Seeking information about a risky medicine: Effects of risk-taking tendency and accountability," *Journal of Applied Social Psychology*, vol. 31, no. 4, pp. 778–795, 2001.

[14] C. Höppner, M. Buchecker, and M. Bründl, "Risk communication and natural hazards," *CapHaz-Net WP5 report*, 2010.

[15] S. Deshan Ilangakoon and K. Y. Abeywardena, "The use of subliminal and supraliminal messages in phishing and spear phishing based social engineering attacks; feasibility study," in *2018 13th International Conference on Computer Science Education (ICCSE)*, 2018, pp. 1–5.

[16] N. Liv and D. Greenbaum, "Deep fakes and memory malleability: False memories in the service of fake news," *AJOB neuroscience*, vol. 11, no. 2, pp. 96–104, 2020.

[17] L. M. Goos and G. Ragsdale, "Genomic imprinting and human psychology: Cognition, behavior and pathology," in *Genomic imprinting*. Springer, 2008, pp. 71–88.

[18] M.-H. Huang, R. Rust, and V. Maksimovic, "The feeling economy: managing in the next generation of artificial intelligence (ai)," *California Management Review*, vol. 61, no. 4, pp. 43–65, 2019.

[19] R. Gruetzemacher, D. Paradice, and K. B. Lee, "Forecasting extreme labor displacement: A survey of ai practitioners," *Technological Forecasting and Social Change*, vol. 161, p. 120323, 2020.

[20] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human factors*, vol. 39, no. 2, pp. 230–253, 1997.

[21] "MechanicalTurk," https://www.mturk.com/, [Online; accessed 15-October-2020].

[22] "Ethics guidelines for trustworthy AI," https://ec.europa.eu/ digital-single-market/en/ news/ethics-guidelines-trustworthy-ai/, [Online; accessed 03-September-2020].

[23] "Recommendations for a Safe and Ethical Transition," https://ec.europa.eu/info/news/new-recommendations-for-a-safe-and-ethical-transition-towards-driverless-mobility-2020-sep-18, [Online; accessed 03-September-2020].

[24] "The IEEE global initiative on ethics of autonomous and intelligent systems. IEEE Standards Association," https://standards.ieee.org/industry-connections/ec/autonomous-systems.html (2019)., [Online; accessed 09-September-2020].

[25] "Ethics framework," https://www.migarage.ai/ethics-framework/ , [Online; accessed 09-September-2020].

[26] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[27] A. H. Kiran, N. Oudshoorn, and P.-P. Verbeek, "Beyond checklists: toward an ethical-constructive technology assessment," *Journal of responsible innovation*, vol. 2, no. 1, pp. 5–19, 2015.

[28] C. Nebeker, R. J. Bartlett Ellis, and J. Torous, "Development of a decision-making checklist tool to support technology selection in digital health research," *Translational behavioral medicine*, vol. 10, no. 4, pp. 1004–1015, 2020.

[29] P. J. Windley, *Digital Identity: Unmasking identity management architecture (IMA)*. " O'Reilly Media, Inc.", 2005.