Ethical-aware autonomous systems from a social psychological lens

Paola Inverardi[1], Massimiliano Palmiero[1], Patrizio Pelliccione[2] and Massimo Tivoli[1]

[1]Department of Information Engineering, Computer Science and Mathematics (DISIM), University of L'Aquila, Via Vetoio, building 'Alan Turing', L'Aquila, 67100, Italy
[2]GranSasso Science Institute, Viale Francesco Crispi, 7, L'Aquila, 67100, Italy

Abstract

In the digital era the use of autonomous software systems has increased exponentially, carrying with it not only comfort and new opportunities for everyone, but also problems related to technical proficiency, usability, and accessibility. In addition, the incremental use of autonomous systems has posed serious ethical issues, which include those related to data protection and security. In fact, users cannot hold the total control of their digital behaviors and information released while interacting with autonomous systems. Therefore, it is crucial to engineer autonomous systems so to preserve human dignity of the users. One possibility is to empower users with a software layer that preserves their moral preferences when they interact with or use autonomous systems. This paper addresses the issue of adopting an interdisciplinary perspective to design a system that uses an ethical software exoskeleton as a software mediator between the autonomous system and the user. This leads to investigate ethics from a social psychological point of view, considering the combination of morality with personality. Thus, a preliminary study aimed at forming ethical profiles is reported. Finally, the benefits of an approach where the person is put at the center of the digital society are briefly illustrated.

Keywords

Autonomous system, Exoskeleton, Privacy, Ethics, Personality.

1. Setting the context

In the last years, information technology has increasingly influenced human lives at the societal, cultural, economic, and political levels. Besides providing a variety of services for the citizens (e.g., health, financial, educational, job search, tax payment, shopping, and transportation services), the widespread use of software systems, notably by means of mobile devices, has also revolutionized social interactions. Together with numerous advantages in daily life, the large use of software systems has brought up problems that cannot be neglected and are becoming increasingly relevant. One of the most important issues is related to the individuals' privacy protection. Indeed, the danger represented by unauthorized disclosure and improper use of personal data by digital systems has been addressed both technologically [1] and in terms of general data protection regulation (GDPR) [2]. In this vein, the growth of autonomous technology and Artificial Intelligence (AI)-based systems has put forwards new risks in their use that have more general ethical implications impacting on human dignity [3, 4]. Europe is at the forefront of confronting these risks. The Ethics Guidelines for Trustworthy AI of European Commission High-Level Expert Group on AI document [5] recommends that AI-based systems satisfy the following requirements: (i) respect the rule of law; (ii) alignment with agreed

Proceedings of CoPDA2022 - Sixth International Workshop on Cultures of Participation in the Digital Age: AI for Humans or Humans for AI? June 7, 2022, Frascati (RM), Italy

EMAIL: paola.inverardi@univaq.it (P. Inverardi); massimiliano.palmiero@univaq.it (M. Palmiero); patrizio.pelliccione@gssi.it (P. Pelliccione); massimo.tivoli@univaq.it (M. Tivoli).
ORCID: 0000-0001-6734-1318 (P. Inverardi); 0000-0002-2051-0126 (M. Palmiero); 0000-0002-5438-2281 (P. Pelliccione); 0000-0001-9290-1997 (M. Tivoli).

ethical principles and values, including privacy, fairness, and human dignity; (iii) keep the humans in control over the system; and (iv) trustworthiness regardless of any possible error of the system. The recent AI act (a proposed law by the European commission) suggests a risk-based approach to control the indiscriminate diffusion of AI-based systems in Europe.

However, we are still far from "keeping the humans in control over the system" and the interaction among autonomous systems and users is still uneven. Producers of autonomous systems still retain the power and the burden to preserve the individual users' rights. As an example of what we mean with "uneven", just consider the present situation implied by the GDPR adoption when accessing web sites. To be compliant with the regulation, and before letting a user interact, a website needs to obtain consent on the way user's data will be managed by the website and its third parties. The way this initial interaction is carried out by the systems can vary greatly and in general not in favor of making the user's life easier. As a matter of fact, if in principle the user is guaranteed by a law, in practice, this may result in leaving her unprotected because of the difficulties to carry over the interaction on the user side. Moreover, preserving human dignity does not mean to just comply with the law, but it goes beyond the law: "the principle of human dignity, understood as the recognition of the inherent human state of being worthy of respect, must not be violated by 'autonomous' technologies" [6].

This means that we should empower the users, and the behavior of autonomous systems should comply not only with the law, but also with users' moral preferences. In other words, it is necessary a human-centric customization of the behavior of autonomous systems to enable the control of the system actions when these may impact users' ethical preferences. If this is not possible, then, the user can refuse to interact with the system. To this extent the preliminary interactions among the systems and the user shall be transparent and trustable to reach a consensual agreement on how decisions are distributed among them. It is worth noticing that this agreement does not need to be of a contractual form as implied by the GDPR but can be inspired by different (also regulatory) protocols.

1.1. The Exosoul project

As shown in Figure 1, the Exosoul project [7] aims at empowering humans with an automatically generated software exoskeleton, which guarantees data protection and privacy and, more in general, human dignity-based on personal ethics and privacy preferences. The exoskeleton would act as an ethical software mediator that adjusts the system's behavior according to the user's soft ethics (personal preferences), without violating the system's hard ethics (values collectively accepted) [8]. The exoskeleton relies on the users' ethical profile to guarantee a fair interaction between the AI autonomous system and the user.

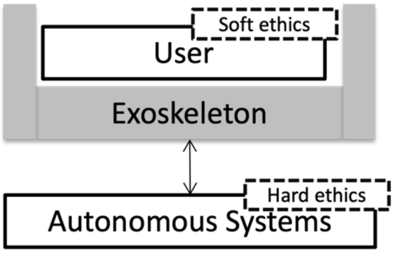

Figure 1Exosoul overview

1.2. Exosoul scenario

In this section, we briefly discuss a possible Exosoul scenario in the automotive and mobile domains, in the setting of a parking lot in the city. The scenario aims to show how the exoskeleton mediates the user-system interaction to adjust (i.e., adapt) the system's behavior based on the user's ethical profile, without violating the system's hard ethics.

Two autonomous connected vehicles (A and B hereafter), with a passenger each, are competing for the same parking lot. Passenger B has a weak-health status. A and B are rented, thus they are multiuser and have a default utilitarian ethics, provided by the cars' producers, which determines their decisions. Thus, the cars will look for the free parking lot that is closer to the point of interest and, in case of contention, the closest car gets in. The passengers interface with the vehicles through their mobile phones, which act on behalf of their owners. A and B are approaching the parking lot. A is closer to the parking lot and would take it. However, by communicating with B, it receives the information that passenger B is in weak-health condition (privacy tradeoff). A has a generosity ethics that manifests in the presence of weak-health people, and, consequently, actions are taken to leave the parking lot to B. This scenario shows how personal privacy is strictly connected to ethics: by disclosing a personal piece of information, the weak-health passenger manifests the utilitarian expectation that surrounding drivers might have generosity ethics.

In this scenario we can consider three different cases depending on the system compliance to regulations and laws, e.g., GDPR. That is: (V1) the system is strictly compliant with the GDPR and will not manage private information; (V2) the system is compliant with the GDPR, although flexible, i.e., if authorized by the vehicle occupants, it handles private information only until the vehicle does not exit the parking lot; (V3) the system is not compliant with the GDPR, e.g., outside EU, and does not guarantee to destroy private information. In V1, there is no possible mediator and, hence, no exoskeleton to be built. In V2, a mediator can be built to request the destruction of personal data upon exiting the parking lot. In V3, the passenger of B does not disclose the information since the system does not guarantee her privacy.

2. Exploration of ethics

Defining the interaction between autonomous systems and users is not only a computer scientist issue, rather it involves different fields of human and social sciences. Specifically, the implementation of a human-centric customization of the autonomous system aimed at mediating the interactions between autonomous and AI-based systems and users, based on users' moral preferences, implies a theoretical reflection and an empirical investigation on ethics (or morality), using a social science perspective. Ethics is a branch of moral philosophy that concerns what is morally good or bad and what is morally right or wrong. It deals with many different aspects of human life, involving practical judgments, decision making, and actions. This leads ethics to be rooted also in other disciplines (e.g., politics, economics, anthropology, sociology, and psychology). Consistent with consequentialism models, moral cognition is guided by a harm-based template, that would help people to discern between rightness and wrongness, according to the amount of harm resulting from an action [9]. By contrast, consistent with deontological models, moral cognition depends on conformity to standards and principles socially defined, whose violations are considered wrong [10]. Thus, some individuals consider the consequences of their actions, whereas others adopt actions following principles of justice and fairness.

Notably, one of the most important theories, which is consistent with the consequentialism-deontological dichotomy is the Ethics Position Theory [11], has been used as the basic theoretical framework in our preliminary study. This theory distinguishes between relativism, which relates to people's beliefs about consequences, and idealism, which relates to conformity to absolute principles. Given that individuals can range from high to low in relativism and idealism, the theory identifies four perspectives on morality: (i) *situationism* (high relativism/high idealism), involving a commitment to promoting human well-being; (ii) *subjectivism* (high relativism/low idealism), involving realism, which means no strong endorsement on moral standards and no 'do-not-harm' mandate; (iii) *absolutism* (low relativism/high idealism), relying on moral standards and based on harm minimization; and (iv) *exceptionism* (low relativism/low idealism), based on conventionalism and tolerance of exceptions to moral standards [12].

2.1. The role of personality in eliciting ethical preferences

Ethics requires to deal with many different factors that may affect moral judgments and actions, including the individual's personality [13], which has a clear relapse on moral behavior. Specifically, on the one hand, the traits of honesty/humility (sincerity, fairness, and genuineness in dealing with others) and conscientiousness (orderliness, meeting of obligations, self-discipline, integrity, and fairness) are related to morality and integrity [14] and are also high-resistant to moral disengagement [15]. Consistent with the Ethics Position Theory, these traits are related positively to idealism and negatively to relativism [15], although conscientiousness can also be positively related to relativism [12]. On the other hand, Machiavellianism (lack of moral standards and manipulation), narcissism (feelings of entitlement and ego-inflated behaviors), and psychopathy (cynicism and sadism) [16] affect negatively moral judgments and ethical behaviors, dismissing morality as a guide for actions. Moreover, Machiavellianism, narcissism, and psychopathy are positively related to relativism and negatively to idealism [12].

2.2. A preliminary study

To define users' ethical profiles that are predictive of digital behavior, our research group conducted a preliminary study considering that differences in moral judgments depend on the combination of ethical positions and personality, also including worldviews, which are the person's constructs and assumptions for understanding the world. We administered a questionnaire to 317 young individuals to measure:

- morality in terms of ethics positions (idealism and relativism), personality traits (honesty/humility, conscientiousness, Machiavellianism, and narcissism) and the worldview (normativism); and
- the digital behavior in terms of privacy violation, copyright infringement and caution.

Two clustering approaches were pursued to create ethical profiles that are predictive of digital behaviors. The first clustering approach was aimed at exploring the dataset using the classical two-step method (the explorative hierarchical analysis followed by the confirmative k-means analysis). The second clustering approach was aimed at exploring the extent to which the four moral perspectives (situationism, subjectivism, absolutism, and exceptionism) defined by Ethics Positions Theory could be confirmed by a k-means analysis.

Thus, the first approach based on the dendogram, i.e., the agglomeration schedule coefficients, and the interpretability of the clusters showed a 2-cluster solution (silhouette value = 0.3). Based on the distribution of the input variables between the two clusters, the first was defined 'opportunist' (210 subjects) and the second 'virtuous' (107 subjects). Concerning the prediction of digital behavior, the 'virtuous' cluster scored lower in privacy violation and copyright infringement, and higher in caution than the 'opportunist' cluster. In other words, the 'virtuous' cluster is guided by principled values, whereas the 'opportunist' cluster is more prone to violate rules to achieve personal benefits. In addition, the 'virtuous' cluster pays more attention to privacy setting and information provided during registration to Internet websites and services.

The second approach based on an a-priori 4-cluster solution showed that the clusters were well balanced in terms of numerosity of subjects, although some overlap between the input variables across the clusters was found especially in terms of Machiavellianism and Narcissism (silhouette value = 0.2). Based on the distribution of the input variables between the four clusters, the identified clusters are: 'legalist' (86 subjects),'sensible' (79 subjects),'opportunist' (77 subjects) and 'virtuous' (75 subjects). Regarding the prediction of digital behavior, the 'legalist', 'virtuous', and 'sensible' clusters showed no difference in privacy violation whereas the 'opportunist' cluster showed higher scores than the other 3 clusters. In addition, the 'legalist' cluster showed no difference as compared to the 'sensible' and 'virtuous' clusters, but lower scores than the 'opportunist' cluster, which showed higher scores than the other 3 clusters in terms of copyright infringements. Finally, the 'virtuous' cluster showed higher scores than the 'opportunist' cluster in terms of caution.

In general, the 2-cluster solution appears more actionable than the 4-cluster solution. Indeed, the 2-clueter solution provides reliable results not only as a function of the input variables reflecting ethics, personality, and normative worldview (less overlap between the clusters, higher silhouette value), but also allows to better predict digital behavior than the 4-cluster solution. In conclusion, this study has shown that ethical profiles can be conceived as a set of factors, including bright and dark personality traits. Further refinements of ethical profiles are in progress (adding more factors) to increase the reliability of the variety of users' moral preferences.

3. Conclusions

In a digital society where human beings and autonomous machines continuously interact, there is a need to redefine the interaction, use, and collaboration with autonomous machines. Our approach takes an ethical perspective and proposes to adjust the behavior of autonomous systems according to user's moral preferences. This necessitates of a multidisciplinary expertise that also includes the study of people's ethical thinking: only by exploring ethics in its multifaceted aspects it is possible to deeply understand humans' moral thought and behavior and subsequently to implement an exoskeleton that protects the human dignity of individuals while interacting with autonomous systems. Therefore, by combining computer science with social science, we propose to empower humans by reflecting their moral preferences in their digital interactions. This implements a paradigm shift from a static to a ground-breaking dynamic approach for ethical-aware software systems. That is, the exoskeleton embodying the user's ethical profiles assists users in their digital behaviors exploiting their own moral preferences and molds the adaptation of the autonomous system. Adjusting the behavior of autonomous and AI-based systems can increase humans' awareness and expression of ethics, privacy, and dignity, and ultimately promote a digital world where all parties are in better balance of power.

4. Acknowledgements

We thank the Exosoul team for discussions and inspirations.

5. References

[1] P. Samarati and S. C. de Vimercati. Access control: Policies, models, and mechanisms. In R. Focardi and R. Gorrieri, editors, Foundations of Security Analysis and Design, pages 137–196, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[2] The European Parliament and the Council of the European Union. Eu general data protection regulation (gdpr) - regulation eu 2016/679 of the european parliament and of the council of 27 april 2016. Official Journal of the European Union, 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[3] P. Inverardi, The european perspective on responsible computing, Communications of the ACM 62 (2019) 64–64. doi:10.1145/3311783.

[4] P. Inverardi. The challenge of human dignity in the era of autonomous systems, in: H. Werthner, E. Prem, E.A. LeeCarlo Ghezzi (Ed.), Perspectives on Digital Humanism, Springer, Cham,2022, pp. 25–29 doi:10.1007/978-3-030-86144-5_4.

[5] High-Level Expert Group on AI. The Ethics Guidelines for Trustworthy Artificial Intelligence, 2019. URL: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

[6] European Group on Ethics in Science Statement on artificial intelligence, robotics and 'autonomous' systems. shorturl.at/mzBR8, 2018. URL: https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.

[7] M. Autili, D. Di Ruscio, P. Inverardi, P. Pelliccione, M. Tivoli, A software exoskeleton to protect and support citizen's ethics and privacy in the digital world, IEEE Access 7 (2019) 62011–62021. doi:10.1109/ACCESS.2019.2916203.

[8] L. Floridi, Soft ethics and the governance of the digital. Philosophy & Technology 31 (2018) 1-8 doi: 10.1007/s13347-018-0303-9.

[9] K. Gray, L. Young, A. Waytz, Mind perception is the essence of morality, Psychological Inquiry 23 (2012) 101–124. doi:10.1080/1047840X.2012.651387.

[10] R. Janoff-Bulman, N.C. Carnes, Surveying the moral landscape: Moral motives and group-based moralities, Personality and Social Psychology Review 17 (2013) 219–236. doi:10.1177/1088868313480274.

[11] D.R. Forsyth, Judging the morality of business practices: The influence of personal moral philosophies, Journal of Business Ethics 11 (1992) 461–470. doi:10.1007/BF00870557

[12] D.R. Forsyth, Making moral judgments: Psychological perspectives on morality, ethics, and decision-making. Routledge, New York, NY, 2019.

[13] L.J. Walker, J.A. Frimer, Moral personality of brave and caring exemplars, Journal of Personality and Social Psychology 93(2007) 845–860. doi:10.1037/0022-3514.93.5.845.

[14] T.R. Cohen, A.T. Panter, N. Turan, L. Morse, Y. Kim, Agreement and similarity in self-other perceptions of moral character, Journal of Research in Personality 47 (2013) 816–830. doi:10.1016/j.jrp.2013.08.009.

[15] B.T. Ogunfowora, V.Q. Nguyen, P. Steel, C.C. Hwang, A meta-analytic investigation of the antecedents, theoretical correlates, and consequences of moral disengagement at work, Journal of Applied Psychology. Advance online publication (2021). doi:10.1037/apl0000912.

[16] D.L. Paulhus, K.M. Williams, The dark triad of personality: Narcissism, Machiavellianism, and psychopathy, Journal of research in personality 36 (2002) 556–563. doi:10.1016/S0092-6566(02)00505-6.