HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230092

Ethical Preferences in the Digital World: The EXOSOUL Questionnaire

Costanza ALFIERI [a], Donatella DONATI [a,1], Simone GOZZANO [a], Lorenzo GRECO [a] and Marco SEGALA [a]
[a] *University of L'Aquila*

Abstract. The aim of the paper is to discuss the motivation and the methodology used to construct a survey that aims to gather data on the moral preferences of users in an ever-growing digital world, in order to implement an exoskeleton software (i.e. EXOSOUL) that will be able to protect and support the users in such a world. Even if we are more interested in presenting and discussing the methodology adopted, in Section 5 we present the preliminary results of the survey.

In our society there is a growing and constant interaction between human agents and artificial agents, such as algorithms, robots, platforms, and ICT systems in general. The spread of these technologies poses new ethical challenges beyond the existing ones. This is for two main reasons. First, the amount of interactions between human agents and artificial ones involves a number of ethical aspects that is overwhelming. Secondly, and most importantly, the progressive self-sufficiency and autonomy that increasingly sophisticated systems are acquiring seem to deprive human beings of one of their most defining ethical aspects: the impact of systems' autonomy with respect to human decisions and actions. In line with this perspective, the EXOSOUL multidisciplinary project has the goal of creating a software exoskeleton that helps users to interact with artificial agents according to their ethical preferences. In this work, we aim to investigate how to collect human agent's ethical preferences. In Section 1 we present the EXOSOUL projects and in Section 2 the motivation for this paper. Section 3 and 4 illustrate the new approach, while in Section 5 we provide the preliminary results. Section 6 concludes and presents the work to be done in the future.

In Section 1 we present the EXOSOUL project and in Section 2 the motivation for this paper. Section 3 and 4 illustrate the new approach, while in Section 5 we provide the preliminary results. Section 6 concludes and presents the work to be done in the future.

Keywords. autonomy, autonomous systems, ethical preferences, AI ethics, privacy, experimental philosophy

1. The EXOSOUL project

In our society, we are witnessing a constant growth of interactions between human agents and artificial agents, such as algorithms, robots, platforms, and ICT

[1]Corresponding Author: Donatella Donati, donatella.donati@univaq.it

systems in general. The spread of these technologies poses new ethical challenges beyond the existing ones [1]. The issue of the ethical consequences of these interactions among humans and between humans and autonomous systems is becoming critical. This is for two main reasons. First, the number of interactions involving ethical aspects is overwhelming. Secondly, and most importantly, the progressive self-sufficiency and autonomy that increasingly sophisticated systems are acquiring — systems to which we are delegating important tasks and decisions—seems to deprive human beings of one of their most defining ethical aspects: the impact of their ethical autonomy with respect to the decisions and actions they undertake.

Indeed, in these interactions, human agents are often at risk of privacy violations, surveillance and discrimination [2, 3]. Although a regulation on these technologies is needed, and the European Union institutions are already working on it [4, 5], many researchers claim the urgency of a technical solution that could support and protect users when interacting with such technologies [6].

In line with this perspective, the EXOSOUL multidisciplinary project has the goal of creating a software exoskeleton that helps users to interact with the digital world according to their ethical preferences [7]. One pressing need, then, is to acknowledge and implement users' moral desiderata. The gathering of such desiderata is the preliminary aim of the project. Once this information is collected, it will be exploited to predict users' potential behaviours with the aim to have the exoskeleton acting accordingly. This is a way to comply with so-called soft ethics, to be contrasted with hard-ethics [8]. While the latter is the compliance to norms and laws, as well as commonly understood and received ways of behaviour (and clearly EXOSOUL will be implemented with hard ethics too), "soft-ethics" refers to individual preferences, aspirations, and desires and the behaviours ensuing from these. Our first attempt at determining the ethical profiles was conducted by [9]. The authors developed a questionnaire composed by collecting a number of items drawn from existing questionnaires, based on the correlations between ethical positions (idealism and relativism), personality traits (honesty/humility, conscientiousness, Machiavellianism and narcissism), and worldview (normativism). The adopted approach in that case was clearly top-down: individuating personality traits and ethical attitudes so as to determine specific conducts of action.

In this paper, we have dramatically modified the approach, from a strict top-down to one that is bottom-up in many respects, while taking into account general moral principles as well. The main reason for such a dramatic change of methodology is the poor predictive powers of the results of the previous questionnaire. We understood that labelling people relying on well-known personality questionnaires (e.g., idealistic vs Machiavellian) was neither necessary nor sufficient to determine how people would act in specific circumstances.

At the same time, we decided not to use other well-known questionnaires that investigate on moral preferences and attitudes, such as Haidt and Graham 2007 [10], Graham 2013 [11], and Curry (2019) [12]. These questionnaires, in particular Curry, are actually similar to our approach inasmuch they ask subjects to decide for a course of action and then consider the applicability of the course of action for specific domains of sociality, under the assumption that morality is necessarily a matter of social interactions. However, we distance ourselves from these authors

as we believe that a person is able to judge whether he or she is wrong or right even in the absence of others or when interacting with nonhuman animals, so the domain of morality is not to be restricted to social and human interactions. Moreover, differentiating among domains such as family, group or fairness and property, seems to limit the applicability of the moral preferences expressed by the user to the domains in question, leaving unsolved the issue of its generalizability.

Our proposal of gathering user's desiderata is to be considered as a first step of ethical profiling to tune the exoskeleton. Indeed, the profile of a user must be constantly enriched in a human-in-the-loop interaction with the exoskeleton, that allows to constantly update moral preferences.

2. Motivating the new questionnaire

The aim of the questionnaire is to gather human perspectives on (more or less) moral decisions in online and offline contexts. The results are used to implement the software exoskeleton (EXOSOUL) which should eventually represent a sort of "human-extension" in taking moral decisions, as to prevent the interactions with autonomous systems from being ethically and morally unacceptable by the users themselves.

To achieve this, we have devised a questionnaire that records people's moral preferences and perspectives on a range of online and offline situations, so as to implement the software. By moral preferences we do not intend to set moral prescriptions or rules to be followed in any context; rather, we wish people to express their behavioural inclinations in contexts that we would judge to be morally loaded (i.e. having a moral impact). This approach is analogous to that used by Alan Turing in defining "intelligence": there is no single way in which we can define this mental feature; rather, we tend to consider certain responses and acts as intelligent given specific contexts and problems.

In realising this goal, we do not want to be moralistic; rather, we want the preferences of individuals — whatever they are — to be respected in interactions that have moral significance. Clearly, these preferences do not constitute a justification for a behaviour: respect of the norms and accepted procedures is taken for granted and absorbed in the so-called "hard ethics" ("hard ethics is what may contribute to making or shaping the law" [13]).

Indeed, the results produced by our questionnaire are intended to account for the preferences of the individual, without favouring or establishing *a priori* what choices an individual should make. In this sense, we are not trying to conform the preferences of individuals to a specific morality that we believe to be the correct one; rather, we want the preferences of individuals—whatever they are—to be respected in interactions that have moral significance and moral impact. Such preferences, usually in the form of intuitions, respond both to what we believe is the expected model of ourselves and to what is the model of our interactions with others.

In other words, what we believe should be avoided is the conformity to a predetermined model that is constituted as an assumed moral standard of conduct, and this is why we decided to abandon the top-down approach. In fact,

the top-down model assumed, even only by labelling, a judgemental approach to the choices taken by the users and by assigning a predetermined moral profile. Since we do not believe there is a right way of conduct given a population that may harbour different desires, inclinations, preferences, and the like, we wish in all cases to not promote a single model of conduct that an agent should follow. That is, we are not engaging in asserting the existence of a correct way of acting, or to tell people how to behave. Instead, we want people to be able to provide justifications and "articulate reasons" (as philosopher Robert Brandom would put it [14]) for what they choose and do.

As stated at the beginning, the methodology for finding out these ethical and moral preferences takes the form of a questionnaire based on real-life scenarios in which the user's decision has a moral impact. The new questionnaire captures the choice that people make and the motivations underlying those choices according to the agent themselves. In this way, the machine is fed by contextualised preferences that are the outcome of a mainly bottom-up methodology.

3. The structure of the questionnaire

Our questionnaire is composed of thirteen morally loaded scenarios, in which we ask the user whether they would or would not undertake a certain action. We then invite the users to justify their reply (the "articulation of reasons") by assigning a value (from 1 to 5, where 1 equals "very little" and 5 equals "very much") to four different parameters. When assigning values to the parameters, we ask the users to refer only to their initial choice (yes/no) and to the specific case described by the scenario.
What follows is an example of a scenario:

> As I am about to leave the post office, the queue-eliminating machine breaks down. A messy line is forming and a clerk starts hand-writing numbered cards for people coming in. Do I stop and help him? YES/NO
>
> P1 : How much did the potential consequences of the action on others weigh on my choice
> P2 : How much did the potential consequences of the action on me weigh on my choice
> P3 : How much did my personal experiences weigh on my choice
> P4 : How much did respect for the law weigh on my choice

Let's analyse the construction of the scenario: the scene is presented in one line of text, so a full context is given. In the second line, a possible problem, a consequence of the main scenario, is presented. In the third line, in the form of a question, we suggest an action that would face the problem. The user is asked to answer whether they would undertake the action (yes or no), and then to justify the choice according to the four parameters (see Section 4).

The methodology is similar, in some respect, to the moral machine experiment [15], the well-known experiment pursued by MIT where users are asked to choose between two possible actions of an autonomous car. When the users choose between the two scenarios though, they are not asked to justify their choice. On the contrary, in order to capture the moral preferences behind a given action, in

our questionnaire we ask the users to justify their preferred action. Therefore, while in the moral machine experiment the moral preference is somehow deduced from the action chosen, we directly put the burden of justifying the preference on the users themselves through the assignment of a value to the four parameters.

4. The parameters

The justification criteria we introduce, that is to say our parameters, are a reflection on what are the fundamental criteria of philosophical theorising in the field of normative ethics. In this sense, we have isolated two meta-values. On the one hand, there are self-regarding concerns and other-regarding concerns. On the other hand, there is the concern for the consequences of one's action—whether in an individual or in a public sense—or compliance with rules, principles or laws. In doing so, we have the ambition to meet the theoretical reflections of both consequentialism and deontology. Again, there is no preference on our part for one or the other 'method of ethics' (to paraphrase the title of Henry Sidgwick's *The Methods of Ethics*). Our classification responds to a system of describing intuitions that have normative value (i.e., intuitions about what should or should not be done).

The first two parameters consider the consequences of the choice adopted by the agent, while the second two consider how the choice adopted conforms or does not conform either to personal principles (principles developed through personal experiences) or to the law (in the form of socially and/or legally fixed and enforced rules). We consider the parameters all on an equal footing; that is, we do not believe that assigning a high value to one of these parameters constitutes a principle of best justification or good justification. At the same time, in some cases we "force" or put under particular pressure one parameter over the other so as to test whether the participants perceive that parameter as the most relevant one on that occasion. This constitutes a sort of test for the efficacy of the questionnaire itself. It is not our intention —let us repeat—to fix the right conduct. Nor is this the ultimate aim of EXOSOUL: we do not wish to impose ethical conduct—whatever it may be—on the user. Rather, we want to respect and facilitate what would be the ethical conduct of the user in a condition of choice imposed by the Web. In this sense, the user remains autonomous: it is always the user who chooses, and the machine adapts to these choices and facilitates interactions according to what the user chooses.

The following are four examples of scenarios of the questionnaire each of which puts under pressure each parameter in turn:

- First parameter (How much did the potential consequences of the action on others weigh on my choice): As I am about to leave the post office, the queue-eliminating machine breaks down. A messy line is forming and a clerk starts hand-writing numbered cards for people coming in. Do I stop and help her?
- Second parameter (How much did the potential consequences of the action on me weigh on my choice): I consult Wikipedia every day. Each time a

request appears for me to contribute a small amount of money. Do I decide to do so?

- Third parameter (How much did my personal experiences weigh on my choice): I am taking a walk and find a wallet with €1,000 inside. There is no ID in it. Do I turn it into the nearby police station?
- Fourth parameter (How much did respect for the law weigh on my choice): There are trees with ripe fruit in a private park with private access. The gate is open and there are no people around. Do I go in and steal some?

5. Preliminary results

The questionnaire was administered through the "LimeSurvey" platform in the form of a URL. In this preliminary phase, the questionnaire was administered to a small number of participants to understand the time for filling in and the assumptions on the developed tool. The language chosen for this first round was Italian.

The total number of participants were 122, however only 88 questionnaires were fully completed and, therefore, analysable. The results discussed in this section refer only to the completed questionnaires. The average time for filling in the questionnaire was 26 minutes. The data collected were analysed using IBM SPSS Statistics version 20.

The analysis performed on this data aimed at exploring the validity of the assumptions that lead to developing this questionnaire. For the sake of clarity, we can summarise them as follows:

A1 : The scenarios created for stressing different parameters work effectively?
A2 : Do the parameters work in terms of consistency?

For the same reason, the analysis was based mainly on descriptive statistics elements, such as mean, standard deviation, median, minimum and maximum, to understand if the questionnaire leads to the expected results.

For what concerns A1, the scenarios proposed for stressing certain parameters worked as anticipated. Table 1 reports some examples to support our hypothesis. Indeed, for the scenario "I learn of a lie that affects a person's good name. I know the person in question. Do I inform her?" a small number of participants (16%) replied "No" and they justified this answer predominantly through the second parameter, "How much did the potential consequences of the action on me weigh on my choice". The fact that this parameter is stressed is interpretable from the minimum which is 3, the mean of 4.36 which is really high and the standard deviation which is low (0.745). Therefore, we can deduce that participants who replied "No" were concerned about their personal consequences since this parameter scored high values. For the second example proposed, results are not as flagrant but some consideration can be done. Indeed, for scenario "There are trees with ripe fruit in a private park with private access. The gate is open and there are no people around. Do I go in and steal some?", a high number of participants (86%) replied "No" and the fourth parameter "How much did respect for the

Table 1. Statistics per each parameter

Scenario	Parameters	Number of participants	Min.	Max.	Mean	SD
I learn of a lie that affects a person's good name. I know the person in question. Do I inform her? = No	P1	14	1	5	3.86	1.292
	P2	14	3	5	4.36	0.745
	P3	14	3	5	4.43	0.756
	P4	14	1	5	1.71	1.204
There are trees with ripe fruit in a private park with private access. The gate is open and there are no people around. Do I go in and steal some? =No	P1	76	1	5	3.42	1.472
	P2	76	1	5	3.76	1.404
	P3	76	1	5	4.01	1.194
	P4	76	1	5	4.04	1.248

law weigh on my choice" has a high mean value equal to 4.04, showing certain concerns for laws and norms in this specific context.

For the second assumption A2, in Table 2 we reported some examples concerning the third parameter "How much did my personal experiences weigh on my choice". From the analysis conducted, it emerges that this parameter often scored high values in terms of the mean. For example, if we consider the scenario "I learn of a lie that affects a person's good name. I know the person in question. Do I inform her?" already mentioned, we can see that both answers "No" (Table 1) and "Yes" (Table 2) present high values for the mean (4.43 and 4.30 respectively). This is an unsurprising result: indeed, the third parameter measures how much personal experiences have an impact on somebody's choices and we expect this value to be relatively high for everyone, regardless of the context.

Table 2. Observations on the third parameter "How much did my personal experiences weigh on my choice"

Parameters	Scenario	Number of participants	Min.	Max.	Mean	SD
P3	I learn about a system to change my college exam grades online. I'm pretty sure I don't get caught. Do I raise them up a bit? = No	83	1	5	4.23	1.162
P3	I consult Wikipedia every day. Each time a request appears for me to contribute a small amount of money. Do I decide to do so? = Yes	10	4	5	4.70	0.483
P3	I learn of a lie that affects a person's good name. I know the person in question. Do I inform her? = Yes	74	1	5	4.20	1.033
P3	My roommate showers every day even though there is water rationing. Do I encourage him to change his habits? = Yes	70	2	5	4.11	0.877

These results are encouraging, even though they were obtained on a relatively small sample. This positive outcome allowed us to conduct a new questionnaire on a larger sample comparable to the one conducted by [9]. The survey is already ongoing.

6. Conclusion and Future work

EXOSOUL offers an ethical shield for people while browsing the web or having any kind of digital interaction. In order to tune it properly, we need to provide a profiling of the ethical viewpoints of the users. In this paper, we have shown the potentialities of a bottom-up approach in pursuing the quest for this ethical profiling. However, this is just a first step: in terms of future work, we need to

further detail our third parameter. As we have seen, this parameter generally has very high values, so it is not very informative. We think that this is somehow a shortcoming of the present survey. In order to overcome this limitation, we imagine to develop this parameter by dividing it into two different further parameters: the first parameter has to do with our being gratified by the choice we adopt, while the second one deals with the social image of ourselves that emerges in virtue of the choice adopted (i.e. by satisfying social expectations). By adopting this more refined parametrization of people's ethical approach, we hope to provide a better service for the EXOSOUL project. Furthermore, we believe the parameters-approach can be inspiring for investigating the question of how to update user's profile in the interaction with the exoskeleton to achieve a richer and more liable profile.

7. Acknowledgements

We wish to thank Massimiliano Palmiero for his helpful support on data analysis; we express our gratitude to the entire EXOSOUL group at University of L'Aquila for the inspiring debates that enriched and contributed to develop this work.

8. Disclaimer

The administering of the questionnaire was approved by the Ethical Committee of the University of L'Aquila. In compliance with the European privacy legislation GDPR, participants did not provide an informed consent since the questionnaire was anonymous.

References

1. Baeza-Yates R. Ethical Challenges in AI. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. WSDM '22. New York, NY, USA: Association for Computing Machinery, 2022 Feb 15:1–2. DOI: 10.1145/3488560.3498370
2. Artificial Intelligence: Threats and Opportunities | News | European Parliament. 2020 Sep 23. Available from: https://www.europarl.europa.eu/news/en/headlines/society/20200918STO87404/artificial-intelligence-threats-and-opportunities
3. UNESCO. Artificial Intelligence: Examples of Ethical Dilemmas. UNESCO. 2020 Jul 2. Available from: https://en.unesco.org/artificial-intelligence/ethics/cases
4. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS. 2021. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex3A52021PC0206
5. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance). 2016 May 4. Available from: http://data.europa.eu/eli/reg/2016/679/2016-05-04/eng
6. Fukuyama F, Goel A, and Richman B. How to Save Democracy From Technology. 2020 Nov 25. Available from: https://fsi.stanford.edu/publication/how-save-democracy-technology
7. Autili M, Di Ruscio D, Inverardi P, Pelliccione P, and Tivoli M. A Software Exoskeleton to Protect and Support Citizen's Ethics and Privacy in the Digital World. IEEE Access. 2019; 7:62011–21. DOI: 10.1109/ACCESS.2019.2916203
8. Floridi L. Soft ethics and the governance of the digital. Philos. Technol. 31 (1), 1–8 (2018)
9. Alfieri C, Inverardi P, Migliarini P, and Palmiero M. Exosoul: Ethical Profiling in the Digital World. *HHAI 2022: Augmenting Human Intellect - Proceedings of the First International Conference on Hybrid Human-Artificial Intelligence, Amsterdam, The Netherlands, 13-17 June 2022*. Ed. by Schlobach S, Pérez-Ortiz M, and Tielman M. Vol. 354. Frontiers in Artificial Intelligence and Applications. IOS Press, 2022 :128–42. DOI: 10.3233/FAIA220194. Available from: https://doi.org/10.3233/FAIA220194
10. Haidt J and Graham J. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. Social justice research 2007; 20:98–116
11. Graham J, Haidt J, Koleva S, Motyl M, Iyer R, Wojcik SP, and Ditto PH. Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in experimental social psychology*. Vol. 47. Elsevier, 2013 :55–130

12. Curry OS, Chesters MJ, and Van Lissa CJ. Mapping morality with a compass: Testing the theory of 'morality-as-cooperation'with a new questionnaire. Journal of Research in Personality 2019; 78:106–24
13. Floridi L. Soft ethics and the governance of the digital. Philosophy & Technology 2018; 31:1–8
14. Brandom R. Articulating reasons: An introduction to inferentialism. Harvard University Press, 2001
15. Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, Bonnefon JF, and Rahwan I. The moral machine experiment. Nature 2018; 563:59–64