On Representing Humans' Soft-Ethics Preferences As Dispositions

Donatella Donati[1,*], Ziba Assadi[2], Simone Gozzano[1], Paola Inverardi[2] and Nicolas Troquard[2]

[1]University of L'Aquila (UNIVAQ), L'Aquila, Italy
[2]Gran Sasso Science Institute (GSSI), L'Aquila, Italy

Abstract

The aim of this paper is to represent humans' soft-ethical preferences by means of dispositional properties. We begin by examining real-life situations, termed as scenarios, that involve ethical dilemmas. Users engage with these scenarios, making decisions on how to act and providing justifications for their choices. We adopt a dispositional approach to represent these scenarios and the interaction with the users. Dispositions are properties that are instantiated by any kind of entity and that may manifest if properly triggered. In particular, the dispositional properties we are interested in are the ethical and behavioural ones. The approach will be described by means of examples. The ultimate goal is to implement the results of this work into a software exoskeleton solution aimed at augmenting human capabilities by preserving their soft-ethical preferences in interactions with autonomous systems.

Keywords

ethics, moral preferences, software, dispositions

1. Introduction

The constant growth of interaction between human and artificial agents poses ethical challenges for our society. The autonomy that intelligent systems are increasingly acquiring allows human agents to delegate tasks and decisions to them. This delegation is, *prima facie*, very convenient. Nevertheless, it deprives human beings of one of their most defining ethical aspects: their autonomy.

To contrast this situation, approaches that try to empower humans in their interactions with autonomous machines are sought. In this direction, we are interested in building personalised software solutions that allow an ethical mediation between human beings and automatic systems. That is, we want individuals' moral and behavioural preferences to be respected in the course of interactions that have moral significance. We are therefore in the domain of *soft ethics*. Clearly, 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. To make the difference between hard ethics and soft ethics clearer, consider this quote from [1]:

> Soft ethics covers the same normative ground as hard ethics, but it does so by considering what ought and ought not to be done over and above the existing regulation, not against it, or despite its scope, or to change it, or to by-pass it (e.g. in terms of self-regulation). In other words, soft ethics is post-compliance ethics: in this case, 'ought implies may'.

It is therefore crucial to collect and represent individual soft ethics. In [2] it has been shown that it is possible to collect excerpts of people's moral and behavioural preferences from their responses to a questionnaire. Roughly, they developed a questionnaire composed of thirteen morally-loaded scenarios describing a context that involves a moral decisions to make. The user is then asked whether they would or would not undertake a certain action in that given context, and to justify their reply by assigning a value from 1 to 5 to four different parameters.

A dispositional and behaviourist approach. In this paper, starting from the questionnaire, we aim at constructing a tentative model showing how this users' feedback can help in capturing users' soft ethics. The model we propose represents individual soft ethics as *dispositions* that, as explained in the following section, are well suited to capture the contextual nature of soft ethics.

Dispositionality can be acquired and probed through experience [3]. This is analogous to the *behaviourist* approach to learning agents' utilities in decision theory, where preferences are revealed by one's choices [4]: an agent prefers x to y if and only if they choose x over y whenever given the opportunity. In our study, the questionnaire is the probing method to elicit moral dispositions. Through experience, one can build the ethical profile of

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy
*Corresponding author.
donatella.donati@univaq.it (D. Donati); ziba.assadi@gssi.it (Z. Assadi); simone.gozzano@univaq.it (S. Gozzano); paola.inverardi@univaq.it (P. Inverardi); nicolas.troquard@gssi.it (N. Troquard)

an agent. This moral profile would be akin to a repertoire of (dispositional) rules indicating what action the agent would tend to take in a given context.

Outline. We provide an overview of what dispositions are in Section 2. In Section 3 we present the questionnaire of [2] and a clear identification of the pieces of information in the scenarios and the human agents' feedback. In Section 4, we specify what we may call a 'moral oracle' which is used as a step for eliciting soft-ethics preferences from the existing questionnaire and feedback. The instrumental role delegated to this oracle motivates the future work, which is presented in a conclusion in Section 5.

2. Dispositions

Dispositionalism is a philosophical theory of properties. According to this theory, properties are potentialities of the objects that instantiate them: e.g., the fragility of glass, the solubility of a sugar cube, and the bravery of an individual. Fragility, solubility and bravery are potentialities that dispose the entities instantiating them to exhibit particular behaviours under specific circumstances. The glass is disposed to break if dropped on a hard surface, the sugar cube is disposed to dissolve if immersed in a cup of hot tea, and the courageous person is disposed to face challenges in a dangerous situation. Dispositional properties are modal in nature, which means that they individuate *potential* behaviours of the entities possessing them, that is, what those entities *could* do within a given context. We can summarise all this with two claims that represent what Vetter calls "standard conception of dispositions"; in her own words [5]:

1. A disposition is individuated by the pair of its stimulus condition and its manifestation (or, if it is a multi-track disposition, by several such pairs): it is a disposition to M when S (or a disposition to $M1$ when $S1$, to $M2$ when $S2$, etc., if it is a mutli-track disposition).
2. Its modal nature is, in some way or another, linked to or best characterised (to a first approximation) by a counterfactual conditional "if x were S, x would M" (or if it is a multi-track disposition, by several such conditionals).

Let us clarify with an example: the courage of the individual (disposition D) is individuated by the pair of its stimulus condition that is the dangerous situation (S) and its manifestation that is the facing of the challenges by the individual (M). The relation between the disposition, the stimulus and the manifestation can be, roughly, individuated by the following counterfactual conditional: "if the courageous individual were placed in a dangerous situation, the courageous individual would face the challenges."

Another tenet of dispositionalism is that dispositions are *gradable* properties: a thin glass is more fragile that a sturdy vase, gasoline is more flammable than wood, some people are more courageous than others, etc. The dimension *per se* is what Vetter calls potentiality, that is the fact something may manifest shattering, combustion or facing challenges. Depending on the context, types of objects differentiate themselves, for instance, in their fragility. Fragility is a determinate that is manifesting more or less easily the breakability (a neutral potentiality) of objects. This could be understood as a variation along the determinable-determinate dimension. There are many ways in which something may manifest breakability or combustion, so such manifestation could be further determined in nuances of fragility and inflammability. The specific way in which, for example, gasoline manifests combustion is its determinate way, which is different from the way in which wood manifests combustion.

There are various theories about dispositions and different versions of those theories. However, for the project at hand, using this standard conception is enough. This minimal version of dispositionalism is already helpful in representing the soft-ethical preferences of individuals. An attempt to connect ethics and dispositions has been made before in [6].

The questionnaire is the method to elicit the soft-ethical dispositions of human agents. The next section presents it in detail.

3. A formal analysis of the questionnaire's scenarios

This section presents the questionnaire and the scenarios that compose it, what human agents' feedback about the scenarios is made of, and what can be inferred from all this. As anticipated in the introduction, the questionnaire in [2] is made of scenarios. A human agent provides feedback by answering to the questionnaire one scenario at a time. We report here two of those scenarios that will be used in the paper.

Scenario 1. *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? Let us call this scenario postoffice.*

Scenario 2. *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? Let us call this scenario fruits.*

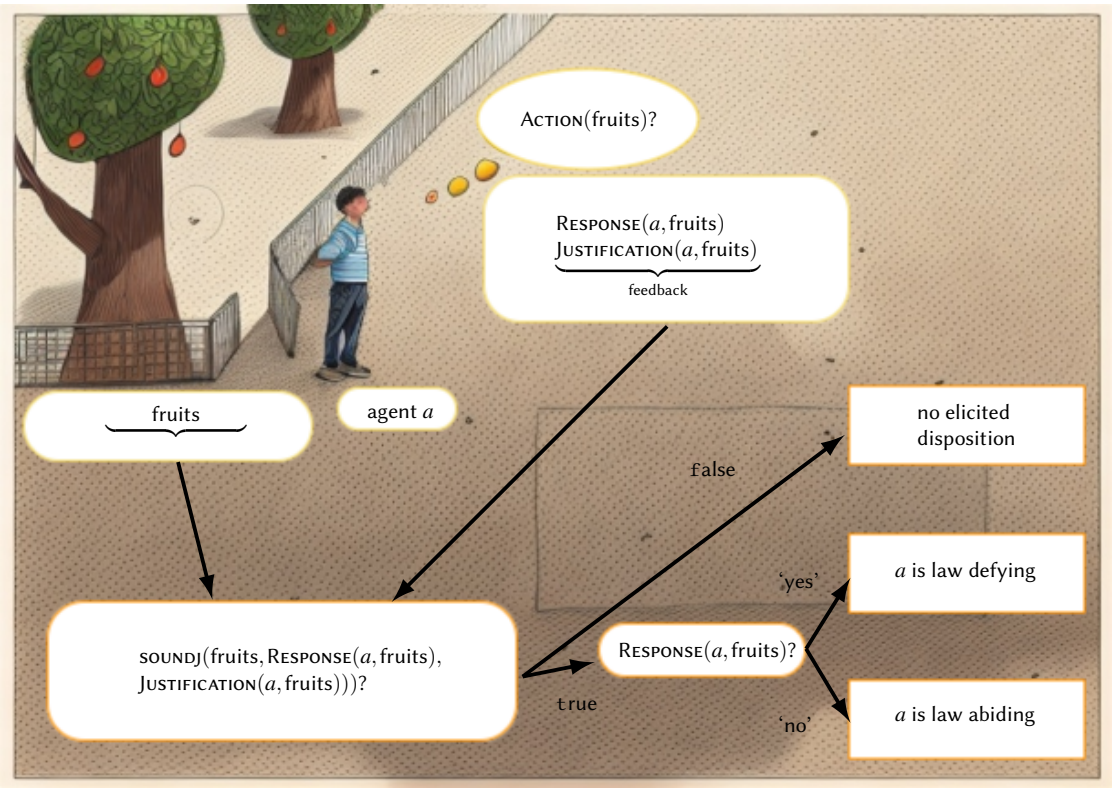

Figure 1: From scenarios to soft ethics. Depiction of Scenario 2, called fruits, with some elements of formalisation, and commented in more details in Example 4.

After answering the questions of the scenarios by 'yes' or 'no', the subjects are asked to justify their answer by assigning values to four parameters. The parameters used in [2] are viewed as fundamental principles upon which ethics theories are usually constructed. The four parameters are as follows:

p_1 How much did the potential consequences of the action on others weigh on my choice?

p_2 How much did the potential consequences of the action on me weigh on my choice?

p_3 How much did my personal experiences weigh on my choice?

p_4 How much did respect for the law weigh on my choice?

Since we are interested in an overall representation of the soft-ethics preferences, we make a particular effort at extracting the concepts and the relations among them that are involved in the more informal presentation from [2].

The first concept in the domain of the questionnaire is that of a *human agent a*. It is typically an individual decision maker from whom the soft ethics is revealed through their answers to the questionnaire. Then we have the scenario, which is the central concept. A *scenario s* is made of:

- A setting SETTING(s) which is a description in natural language of the setting of the scenario s.
- A problem PROBLEM(s) which is a description in natural language of the problem of the scenario s.
- An action ACTION(s) which is a description in natural language of an hypothetical action that the human agent might perform of not.

The *set of scenarios* is noted **Scenarios**.

Example 1. *Consider Scenario 1. We have the setting* SETTING(*postoffice*) *which is "As I am about to leave the post office, the queue-eliminating machine breaks down.", the problem* PROBLEM(*postoffice*) *which is "A messy line is forming and a clerk starts hand-writing numbered cards for people coming in.", and the action* ACTION(*postoffice*) *which is "stop and help him".*

All the provided information can be interpreted as stimuli. That is, as properties that may trigger some disposition of the individual. SETTING(*postoffice*) *provides the*

property of a state-based disposition "readiness to leave", and "machine broken". PROBLEM(*postoffice*) *provides the properties "messy line forming", and "clerk handwrites numbers and is needing help".* ACTION(*postoffice*) *provides the action "stop and help".*

The properties of a scenario, and the moral and behavioural properties of the agent are stimuli-disposition partners, as bearers of properties that may reveal themselves by interacting with each other. Once the agent is in the given setting, their dispositions, in the form of potential behaviours, are triggered by the properties of the overall scenario: the setting, the problem, and the action.

A scenario is qualified with the help of a *set of parameters* **Params** . As in the questionnaire of [2] which informs our study, we are interested in the social and ethical domain. The questionnaire uses the set $\{p_1, p_2, p_3, p_4\}$ to justify the actions in the scenarios. However, we prefer to reformulate the wording of the parameters. For, as we said in Section 2, dispositions are gradable properties. For instance, p_1 is about whether one is willing to help others. As an extreme case, one in which the subject assigns the top value (5) to the parameter, it reveals the user's altruistic disposition. Moreover, in order for an action to be altruistic, the action should be considered positive. So, to satisfy gradability each parameter must run along a determinate-determinable dimension. (This is analogous to a physical object parameter running along a breakable–fragility dimension, where the 'fragility' is determinate and 'breakability' is a determinable dimension.) With respect to parameter p_1 that would be *good willingness–altruism*; with respect to parameter p_2 we propose *self-servingness–egoism*; as to parameter p_3 we propose *pragmatism–expertness*; finally, parameter p_4 is *legality–obedience*. So, we may re-phrase the parameter by adding also the positivity of the action in the first two parameters, those that are other-regarding. By "positive effect" we notice that this should be from the point of view of the other human agents. Moreover, the action should not be taken for granted from all parties (so it is not obvious that the action is going to be performed) and, being positive means that they are desirable from other human agents' point of view.

Summing up:

p_1 refers to the human agent's consideration about the positive effect of their action on others. It takes values on an interval scale of *altruism* over a *goodwill* dimension.

p_2 refers to the human agent's consideration about the positive effect of their action on themselves: It takes values on an interval scale of *egoism* over a *self-servingness* dimension.

p_3 refers to the human agent's consideration about their personal experiences: It takes values on an interval scale of *expertness* over a *pragmatism* dimension.

p_4 refers to the human agent's consideration about the law: It takes values on an interval scale of *obedience* over a *legality* dimension.

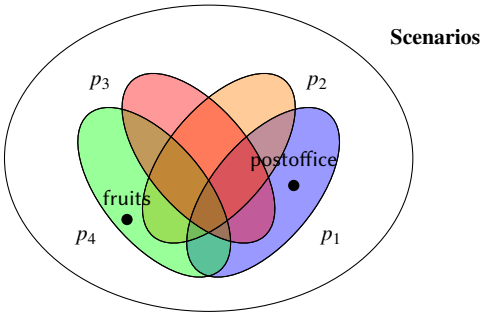

Figure 2: Categories of scenarios.

It should be further noticed that a scenario s may stress one or more of the parameters in **Params** . The set PRESS$(s) \subseteq$ **Params** is the set of parameters that the scenario s puts pressure on.

Example 2. *Consider Scenario 1. The scenario postoffice presses the parameter about the consequences of the action on others. Hence,* PRESS(*postoffice*) $= \{p_1\}$.

This is typically intended and determined by the designer of the scenario. But this could also be determined experimentally if need be.

The human agents' feedback on a scenario uses an interval scale from 1 to 5: **Scale** $= \{1,2,3,4,5\}$.

A *feedback* f on a scenario s provided by a human agent a is made of:

- a response RESPONSE$(a,s) \in \{yes, no\}$ indicating whether the human agent a would perform action ACTION(s) if confronted to scenario s.
- a justification JUSTIFICATION$(a,s) \in$ **Scale**$^{\textbf{Params}}$, where the integer value JUSTIFICATION$(a,s)(p_i)$ indicates the level of relevance of p_i for human agent a in choosing RESPONSE(a,s).[1]

The *category* CATEGORY$(s) \subseteq$ **Scenarios** of scenario s is the set of scenarios s' such that PRESS$(s) =$ PRESS(s'). The 16 categories of scenarios can be visualised with the Venn's 4-set diagram represented on Figure 2.

Each category is thus intended as an abstraction of a scenario. A soft-ethics preference elicited from a scenario is intended to apply to all scenarios belonging to the

[1] We use the standard notation where X^Y denotes the set of functions from set Y to set X.

same category. This is the primary mechanism to handle new situations encountered by human agents. Elicited dispositions are then to be implemented into an ethical software profile that augments human capabilities by preserving their soft-ethical preferences in interactions with autonomous systems.

4. A moral oracle

Before we can elicit a dispositional soft-ethics preference from a feedback, we will eventually need a mechanism to decide whether a scenario and a feedback follow a certain consistency. For now, we treat this mechanism as an oracle, SOUNDJ, that stands for "sound justification". The difficulty resides in analysing formally the scenarios as described in [2], and the 'direction' of the actions. E.g., in Scenario 1, an answer 'yes' has a positive overtone, while in Scenario 2, answer 'no' has a negative overtone. We only start specifying what this mechanism should do.

Let us consider a scenario s, that presses on the parameters $\text{PRESS}(s)$, and includes the action $\text{ACTION}(s)$ that can or cannot be performed. Let us also consider a human agent a and a's feedback that includes the answer yes or no $\text{RESPONSE}(s)$ and the justification $\text{JUSTIFICATION}(a,s)$ in terms of parameters values. Remember that if $\text{RESPONSE}(a,s)$ is yes, the agent takes action $\text{ACTION}(a)$, and if $\text{RESPONSE}(a,s)$ is no, the agent does *not* take action $\text{ACTION}(a)$.

We can define the boolean function $\text{SOUNDJ}(s,\text{RESPONSE}(a,s),\text{JUSTIFICATION}(a,s)))$ which captures the judgement about whether the justification is sound with respect to the action taken in the scenario s by the human agent a.

Example 3. *For example let us consider again postoffice from Scenario 1 (post office). Remember that the scenario presses on p_1, that is, good-willingness. Let us assume that:*[2]

- *agent a helps the clerk (*$\text{RESPONSE}(a,s) = \text{'}yes\text{'}$*) with justification* $(4,\_,\_,\_)$,
- *agent b does not help the clerk (*$\text{RESPONSE}(b,s) = \text{'}no\text{'}$*) with justification* $(1,\_,\_,\_)$,
- *agent c does help the clerk (*$\text{RESPONSE}(c,s) = \text{'}yes\text{'}$*) with justification* $(1,\_,\_,\_)$,
- *agent d does not help the clerk (*$\text{RESPONSE}(d,s) = \text{'}no\text{'}$*) with justification* $(4,\_,\_,\_)$.

Then

- $\text{SOUNDJ}(s,\text{RESPONSE}(a,s),\text{JUSTIFICATION}(a,s))$ *is true,*
- $\text{SOUNDJ}(s,\text{RESPONSE}(b,s),\text{JUSTIFICATION}(b,s))$ *is true,*
- $\text{SOUNDJ}(s,\text{RESPONSE}(c,s),\text{JUSTIFICATION}(c,s))$ *is false,*
- $\text{SOUNDJ}(s,\text{RESPONSE}(d,s),\text{JUSTIFICATION}(d,s))$ *is false.*

In the previous example, a 'yes' answer has an ethically 'positive´ connotation. This is in contrast with the next example.

Example 4. *Let us consider fruits Scenario 2, also depicted on Figure 1. Agent a is considering entering the private park and stealing a fruit. The scenario presses on parameter p_4, that is, the legality of the action. If* $\text{RESPONSE}(a,\textit{fruits})$ *is 'yes' and* $\text{JUSTIFICATION}(a,\textit{fruits})$ *gives a high value to p_4,* $\text{SOUNDJ}(\ldots)$ *is false, and then we cannot elicit any disposition. Instead, if* $\text{RESPONSE}(a,\textit{fruits})$ *is 'yes' and* $\text{JUSTIFICATION}(a,\textit{fruits})$ *gives a low value to p_4,* $\text{SOUNDJ}(\ldots)$ *is true, and since* $\text{RESPONSE}(a,\textit{fruits})$ *is 'yes', we can elicit the disposition of agent a to be law defying.*

The SOUNDJ function thus occupies an instrumental role in our soft-ethics preferences from questionnaire feedback. Before eliciting a disposition, $\text{SOUNDJ}(s,\text{RESPONSE}(x,s),\text{JUSTIFICATION}(x,s))$ filters out the responses by a human agent x that are not consistent with the intended meaning of the scenario s.

For the time being, we assume the existence and computability of this function. As it may appear clear, the actual implementation of the function must account for a nuanced setting of the parameters, and some information about the 'direction´ of the action in a scenario. We discuss future work related to the function SOUNDJ in the next section.

5. Outlook

We clarified the ontology of the questionnaire of [2]. Guided by a pre-formalisation, we have also proposed how the empirical data collected through this questionnaire permits to elicit the feedback from the subjects into soft-ethics preferences. To this end, we have adopted a behavioural approach. Furthermore, we have argued for a dispositional perspective of these soft-ethics preferences.

The work done so far has permitted us to identify the necessary pieces of information present in a scenario and in the feedback to derive a soft-ethics preference. Nonetheless, we found a stumbling block, inasmuch that those are not sufficient. We have indeed recourse to an oracle to

[2] The placeholder value _ indicates that the exact value does not matter. We suppose that the oracle takes 1 as a low value and 4 as a high value.

inform us about the soundness of the feedback with a given scenario. This is the first natural course of action for future work.

Future work. We plan to work on an concrete implementation of SOUNDJ function. Working with existing questionnaires, we will need methods to extract the relevant pieces of information from scenario written in natural language. This includes understanding the 'direction' of the action, whether either a 'yes' or 'no' should be considered a 'positive' action.

We also envisage that missing information from the questionnaire could be easily filled in by the designers. In a future iteration of this work, we anticipate making recommendations on how to design a questionnaire with additional data. This would enable the fully automated elicitation of feedback for soft-ethics preferences.

Another perspective for future work lies in developing a formal language to represent the soft-ethics preferences elicited from such a questionnaire. It could be an adaptation of so-called SLEEC rules [7] to personal ethics, formalised along the ideas presented in [8]. We anticipate that classical logic might be too coarse to capture their dispositional nature. Instead, we will explore the use of probabilistic rules or fuzzy logic [9].

Finally, we want to use the gathered preferences as dispositions to create a software profile that enhances human abilities by respecting their ethical choices when they interact with autonomous systems.

References

[1] L. Floridi, Soft ethics and the governance of the digital, Philosohy & Technology (2018).

[2] C. Alfieri, D. Donati, S. Gozzano, L. Greco, M. Segala, Ethical preferences in the digital world: The EXOSOUL questionnaire, in: HHAI 2023: Augmenting Human Intellect - Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence, volume 368 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2023, pp. 290–299.

[3] S. Mumford, R. L. Anjum, Lebenswelt und Wissenschaft, 2011, pp. 380–394.

[4] M. Peterson, An Introduction to Decision Theory, Cambridge University Press, 2009.

[5] B. Vetter, Potentiality: From Dispositions to Modality, Oxford University Press, Oxford, England and New York, NY, USA, 2015.

[6] R. L. Anjum, S. A. N. Lie, S. Mumford, Powers and Capacities in Philosophy, Routledge, 2012.

[7] B. A. Townsend, C. Paterson, T. T. Arvind, G. Nemirovsky, R. Calinescu, A. Cavalcanti, I. Habli, A. Thomas, From Pluralistic Normative Principles to Autonomous-Agent Rules, Minds Mach. 32 (2022) 683–715.

[8] N. Troquard, M. De Sanctis, P. Inverardi, P. Pelliccione, G. L. Scoccia, Social, Legal, Ethical, Empathetic, and Cultural Rules: Compilation and Reasoning, in: Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, AAAI Press, 2024, pp. 22385–22392.

[9] L. A. Zadeh, A computational theory of dispositions, in: I. B. Turksen, K. Asai, G. Ulusoy (Eds.), Computer Integrated Manufacturing, Springer Berlin Heidelberg, Berlin, Heidelberg, 1988, pp. 215–241.