AI & SOCIETY
https://doi.org/10.1007/s00146-026-02888-z

RESEARCH

Individual ethics and dispositions in the digital world

Donatella Donati[1] · Simone Gozzano[1] · Paola Inverardi[2] · Nicolas Troquard[2]

Received: 9 May 2025 / Accepted: 20 January 2026

Abstract
Personal ethical preferences are key enablers for the design of autonomous systems that respect humans' moral rights and values. This goes beyond embedding ethical and legal principles in the design of the system once and for all. It requires the ability to elicit personal soft ethical preferences, represent them in a digitally useable format, and link them to the individual for use when interacting with digital systems. The aim of this paper is to represent soft ethical preferences through dispositions. Dispositions are properties that are instantiated by any kind of entity and that may manifest if properly triggered; we will focus on moral dispositions of individuals. The dispositional properties in which we are interested are ethical and behavioural ones. We propose a general and formal model to elicit and handle individual moral preferences. The model allows for the examination of real-life situations involving moral dilemmas. Users engage with these scenarios, respond based on how they would act, and provide justifications for their choices. Their responses are then analysed to identify tendencies toward certain actions, which are represented as dispositions. These dispositions can subsequently serve as the foundation for disposition manifestation mechanisms.

Keywords Software agents · Soft ethics · Dispositions · Formal model

1 Introduction

As humans dwell in a world of autonomous systems powered by AI, it is essential to maintain and protect their moral values (such as dignity and autonomy) to live the 'good life' that is dear to moral philosophers (Vallor 2016; Driver 2022). This expectation cannot be entirely delegated to the system's developers, the ones responsible for designing and building the technology. The autonomous and data-hungry nature of current technologies, compounded by the emergence of generative AI, has sparked significant concerns about the potential dangers these uncontrolled technologies may pose to individuals' fundamental rights (European Commission 2018; Inverardi 2019). Among the remedies suggested by institutional and scientific bodies, in addition to the need to regulate the deployment of such technologies in society, which in Europe culminated in the approval of the AI Act (European Parliament 2024), there is a need to empower the user and develop human centric systems (European Data Protection Supervisor 2015; Autili et al. 2025). This includes delegating personal information sources expressing individual preferences to third-party systems that can be chosen and trusted by the user, or to user proprietary software that can interact with the autonomous systems explicitly, expressing the users moral preferences (Autili et al. 2019; Fukuyama et al. 2021; Boltz et al. 2024).

In this work, we are not directly interested in universal machine ethics (Wallach and Allen 2009). Instead, we aim at developing what may be called an 'individual soft-ethical profile'. That is, we position ourselves in a post-compliance ethics, or "soft ethics" (Floridi 2018). We seek personalised software components that empower users in their interactions with autonomous systems, preserving their preferences in ethically charged scenarios. To this aim, we investigate methodologies that would allow one to elicit these soft ethics preferences in a bottom-up fashion from the user, store them, and use them in scenarios possibly not seen before.

✉ Donatella Donati
donatella.donati@univaq.it

Simone Gozzano
simone.gozzano@univaq.it

Paola Inverardi
paola.inverardi@gssi.it

Nicolas Troquard
nicolas.troquard@gssi.it

[1] University of L'Aquila, L'Aquila, Italy

[2] Gran Sasso Science Institute, L'Aquila, Italy

Published online: 08 February 2026

Much of the discourse on machine ethics revolves around weighty issues, such as the use of autonomous weapons, or cars grappling with moral dilemmas, such as choosing between the safety of the occupants and of a jaywalking little girl (Awad et al. 2018). Our focus lies on the moral considerations concerning systems that will serve as our digital partner even and especially in daily mundane life. These systems will confront decisions on our behalf in the ordinary routines of life. For instance, should they disable ad-blocking plugins on independent online newspapers? Should they propose exchanging an earlier electronic reservation slot at a public service with a tattooed man in a wheelchair? Should they lower the thermostat to save energy? Should they suggest ethically sourced products over cheaper alternatives? It is in this context that we are interested in building personalised software solutions that enable 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.

The personalised solutions we are aiming at are thought to apply across the board. Therefore, they can offer empowerment also for individuals whose ethical values are quite different, or even opposed, to one another. Such an issue reminds the so-called "neutrality" of technology. We think this is a burden that we have to accept. As we take free press to be a value independently on the content that could be published, so we think that empowerment should be set as neutral with respect to its use. However, since such empowerment is in the soft-ethics domain, it is dependent on the quality of the laws and the hard ethics determined by those laws. The more stringent the laws and the hard-ethics, the less leverage one can apply to the soft-ethics empowerment. Therefore, there is a trade-off between the space of possible soft-ethics actions and the normative components (the laws) and its direct ethical consequences (hard-ethics).

Individual ethics for software agents As anticipated, we are 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. Floridi (2018) explains the difference between hard and soft ethics as follows:

> Soft ethics covers the same ethical 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'.

Hard ethics is typically implemented by the autonomous system that delivers a given service, e.g., compliance with GDPR for a train ticket reservation web site. Soft ethics is instead individual and implemented in our ethically aware digital personal assistant that is interacting with the system on our behalf thus allowing for the automatic selection of decarbonisation contribution. In a scenario of a care robot interacting with different human patients, we expect the robot to comply with hard ethics, e.g., guarantee the safety of the patients, but we also expect the robot to adapt its behaviour to the different soft ethics of the patients, for example, not forcing her to take a pill should this action injury her human dignity.

A dispositional and behaviourist approach To construct a general model showing how a user's feedback can help in capturing this user's soft ethics, we propose to represent individual soft ethics as *dispositions* that, as explained in Sect. 3, are well-suited to capture the contextual nature of soft ethics.

Dispositions can be acquired and probed through experience (Mumford and Anjum 2011b). This is analogous to the *behaviourist* approach to learning agents' utilities in decision theory, where preferences are revealed by one's choices (Peterson 2009): an agent prefers x to y if and only if they choose x over y whenever given the opportunity. A questionnaire can represent the probing method to elicit moral dispositions, as can be actual experience. One can thus build the moral profile of a human agent. This moral profile would be akin to a repertoire of (dispositional) rules indicating what action the individual would tend to take in a given context.

Dispositions are context-sensitive properties (Choi and Fara 2021). Likewise, the manifestation of a moral disposition depends on numerous factors, including psychological states and the specific contexts in which the person finds themselves. Even slight variations in these contexts can influence how an individual behaves, potentially preventing a particular disposition from manifesting.

Consider a person who is generally known for their kindness. However, in a stressful work environment, where they feel overwhelmed and unsupported, their usual compassionate nature may not surface. Instead, they may react with frustration or detachment. This illustrates how context and psychological factors can impact the manifestation of moral dispositions, which makes them far more nuanced than something like the fragility of a glass, which is more predictable.

We propose a model of ethical dispositions suitable for handling this contextual decision making for software agents.

Outline In Sect. 2, we present related works, ending with the questionnaire of Alfieri et al. (2023) which we take as the basis for the following sections. We provide an overview of dispositions in Sect. 3. In Sect. 4, we present a general dispositional model to elicit soft-ethics preferences from the existing questionnaire and feedback. Section 5 provides a formal definition for the constituent elements of the dispositional model. Section 6 presents disposition elicitation and

manifestation mechanisms. Conclusions and future work are discussed in Sect. 7.

2 Related works

We can divide related works into three broad and overlapping areas: Machine and digital ethics, human values and digital technologies, gathering personal ethics through questionnaires.

Machine and digital ethics The first area of interest focuses on past work on machine ethics that addresses the need to embed "universal" moral principles into autonomous systems. With respect to this kind of work, we are mainly interested in the implementation aspects. In a survey on implementations in machine ethics, Tolmeijer et al. (2021) review the literature along three dimensions: type of ethical theory, non technical aspects, and technical aspects. On the technical dimension, they identify five broad approaches to the implementation of ethical-aware systems: logical rule system, probabilistic engine, statistical learning, optimisation, and case-based reasoning. Our approach will clearly place itself alongside approaches using case-based reasoning. As we will see, however, we will elicit moral dispositions as abstractions of concrete scenarios encountered by the subjects. Although we do not commit to a particular technical implementation, this abstraction must be done using core knowledge and inference in a specific logic. Examples of core knowledge is a person in a wheelchair is a physically impaired person, a fence is a kind access control, etc. More recently, Townsend et al. (2022) proposed a methodology to elicit social, legal, ethical, empathetic, or cultural ('SLEEC') requirements for autonomous systems. Further works on SLEEC rules (Feng et al. 2024; Troquard et al. 2024; Yaman et al. 2025; Townsend et al. 2025) addressed the refinement of high-level principles into lower level operational rules, and the resolution of conflicts and other properties such as redundancy between normative principles.

Digital ethics, as introduced by Floridi (2018), builds upon the work of Floridi and Taddeo (2016) in data ethics. This branch of ethics explores and evaluates moral issues related to data, AI, and digital practices, aiming to promote ethical solutions and establish socially accepted standards that guide digital regulation and governance. In addition, Floridi (2018) distinguishes between hard ethics and soft ethics. Hard ethics is concerned with moral principles that shape, challenge, or influence laws, regulations or even established social norms. It deals with rights, duties, and responsibilities, aiming to establish legal frameworks that enforce ethical standards. A key principle here is that ethical obligations do not always align with legal permissions; sometimes, doing what is morally right means acting against existing laws. In contrast, soft ethics operates within the boundaries of existing regulations, focusing on ethical behaviour beyond legal compliance. It encourages self-regulation and voluntary ethical standards, rather than legal enforcement. In this sense, soft ethics is "post-compliance ethics": what should be done when the law does not require or prohibit specific actions. In our understanding, however, soft ethics concerns not only what is morally good or desirable, but also the space of freedom within what is permitted by hard ethics and the law. It includes the freedom to choose and to shape actions, practices, or values according to individual, institutional, or cultural priorities, as long as they remain within the bounds of legal permissibility.

Human values and digital technologies Another area of interest is the broad area of human values, their elicitation, their application, and their use in the design and implementation of digital technologies, with particular attention to the challenge of value alignment, that is, ensuring that such systems act in accordance with shared human values. Here the literature is vast and multidisciplinary. Human values have been, at least since the end of the last century, taken into consideration in the field of human computer interaction (Friedman and Hendry 2019). In this field, methods to identify relevant stakeholders and to elicit human values have been proposed and distilled (Friedman et al. 2017). More recently the problem of how to embed human values in the design and implementation of software systems, notably the problem of how to operationalise human values has received increasing attention in the software engineering community (Mougouei et al. 2018; Shahin et al. 2022). All of these approaches consider general human-sensitive requirements for the (autonomous) system, which are applicable to groups, categories, and societies of human beings. Closely related is the notion of value alignment, that is, the problem of ensuring that AI systems act in accordance with human or societal values. Gabriel (2020) highlights the normative complexity of this challenge, stressing the need for principled approaches that respect moral pluralism and democratic legitimacy. Nyholm (2023), on the other hand, links value alignment to the issue of responsibility gaps, arguing that meaningful human control is essential to avoid ethical and legal accountability problems in AI systems. Also relevant to our work are actual experiments to elicit and use individual moral preferences. Notably, Awad et al. (2018) present an experiment to uncover the *non personal* 'societal expectations' about the ethical behaviour that autonomous vehicles should follow. Many other experiments have been proposed in the last few years in the AI ethics community, raising also concerns about their validity. In a survey of AI Ethics and preference elicitation, Feffer et al. (2023) suggests evaluating elicitation methods along ten axes. We are not directly concerned with the criticisms of existing experiments as we refrain from entering into the technical details of experimental science. However, we believe that

experiments exploiting our questionnaire model and eliciting moral preferences should follow the recommendations made by the authors.

Gathering personal ethics through questionnaires

Our goal is to construct a model based on individuals' ethical preferences, thereby necessitating a method for their identification and collection. A questionnaire is a widely used survey tool known for its simplicity and effectiveness and it can be particularly useful for understanding and gathering people's ethical preferences. By asking structured questions, one can gain insights into which action individuals would do in a given context. Our aim is to model individuals' soft ethics, which fluctuate based on varying contexts. In doing so, we seek to adopt a descriptivist approach rather than a normative one. Our primary focus is to observe and understand how individuals' moral decisions and behaviours change depending on their moral dispositions and the circumstances, rather than to judge or prescribe what those moral standards should be.

Numerous questionnaires, developed across psychology, sociology, and experimental philosophy literature exist for gathering individuals' ethical preferences, each tailored to specific objectives. Here, we mention a few of them, aiming to clarify why we use one while excluding the other two (which we mention due to their widespread usage).

We first present questionnaires designed for evaluating moral value and related psychological theories. Then, we present two questionnaires designed to elicit what participants would do in hypothetical morally charged scenarios.

Questionnaires for evaluating moral value

One of the most widely used questionnaires for gathering people's moral inclinations is that of Schwartz (1992). The theory delineates ten basic personal values, which in fact form a continuum of motivations for action. Schwartz et al. (2012) refine the set to 19 basic personal values. Benevolence, Conformity, Self-direction, or Hedonism are examples of these values. Schwartz (2012) provides an overview of the personal values and presents two methods to measure them: the Schwartz Value Survey, and the Portrait Values Questionnaire. In our view, Schwartz's theory has a number of shortcomings. Many of the parameters that are used, are expressed in terms which are not purely descriptive. Rather, they possess a normative character, indicating a course of action that we should adhere to, typically emphasising the most positive outcomes in ethical decision-making. However, one complies only with those actions that are in accordance with the values that one endorses. For instance, people tend to assume that benevolence is good and non benevolence is bad, but this depends on the value at stake. So, these parameters tend to conflate moral (the values that we adopt) and ethics (the behaviours and actions that we perform to comply with the adopted values).

Moral foundation theory (Graham et al. 2013) is a theory that views social, cultural, and moral behaviours, where values are universal but contextually variable.

Originally, the theory has five 'foundations' and it emphasises the dyadic opposition between values and violations: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and purity/degradation. The authors allow for a degree of flexibility in the theory, which supports modifications. In Atari et al. (2023), the fairness foundation was split into equality and proportionality.

Moral foundation theory is operationalised in moral foundation questionnaires (MFQ), first presented in Graham et al. (2011) and updated accordingly in Atari et al. (2023). In particular, it makes use of two scales, one for 'relevance' and one for 'judgement'. Relevance asks participants about how relevant are some considerations when deciding on the morality of something. Judgement asks participants how much they agree with a moral statement.

According to morality-as-cooperation (Curry et al. 2019), morality consists of biological and cultural behaviours in response to situations of cooperation that are recurrent in human social life. A questionnaire of the same nature as the MFQ is proposed to test the theory.

Questionnaires for moral action

The Moral Machine experiment was designed by Awad et al. (2018) to gather moral preferences of people to derive *societal expectations* about the moral conduct that autonomous vehicles should follow. The primary objective of the experiment was to identify universal moral principles that could be integrated into autonomous vehicles, enabling them to align their behaviour with human moral inclinations. The ultimate aim was to develop a universal machine ethics. It offers scenarios which are variants of the trolley problem, where an autonomous car faces an unavoidable accident. Users then have to pick between two actions the car could undertake, each with its own moral implications. This questionnaire is more aligned both with our objectives and the method we prefer for discerning individuals' moral preferences. However, it diverges for two main reasons: first, it exclusively targets self-driving cars, and second, its goal is to pinpoint universally accepted moral principles and formulate a universal ethics for machines. Conversely, our focus lies in uncovering personal ethical principles.

The questionnaire presented by Alfieri et al. (2023) corresponds more closely to the requisites of a survey tool for gathering information about personal ethical and behavioural preferences in morally loaded real-life situations.[1] The questionnaire is made of thirteen scenarios. A human agent provides feedback by responding to the questionnaire,

[1] The original model was proposed and developed in Gozzano (1997).

addressing one scenario at a time. Following a choice between performing or not performing a given action, users can justify their decision by assigning a value from 1 to 5 to four distinct parameters. Here are two of those scenarios that will also be used as examples in the present 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 grab one? Let us call this scenario* fruits.[2]

Although we believe that this approach is promising, some limitations remain to exploit the answers to elicit the behavioural tendencies that are adequate for recommending (or autonomously manifest) a certain behaviour in situations not seen before. First, it may be argued that the set of parameters used for characterising a situation is too coarse. Second, the description of scenarios is missing the crucial information about when an answer should be considered consistent, and individuating a disposition. Specifically, it appears crucial to have access to a formal characterisation of the scenarios and the actions. These characterisations may remain hidden from the subject, but they are central for a prospective computational use. Finally, some of the words used to describe the scenarios are "biased" ("messy line", "steal", etc.). Nonetheless, we believe that it is useful for the purposes of our paper and for a first attempt at formalising individual's soft ethics preferences.

3 Dispositions

Dispositionalism is a philosophical theory of properties. According to this theory, properties are dispositions (i.e., causal powers) of the entities that instantiate them: e.g., the *fragility* of a glass, the *solubility* of a sugar cube, and the *bravery* of an individual. Fragility, solubility and bravery are properties that dispose the entities instantiating them to exhibit particular behaviours under specific circumstances (i.e., in certain contexts). 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, they individuate *potential* behaviours of the entities possessing them. To put it simply: such behaviour does not have to be necessarily manifested by the entity in question, dispositions only individuate what 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 (Vetter 2015):

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 multi-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).

Indeed, dispositional statements are typically connected to subjunctive conditionals. In much of the philosophical literature, dispositions have been analysed by means of conditionals in which the antecedent specifies the triggering stimulus conditions, and the consequent describes the manifestation. The most basic version of this approach is the so-called *Simple Conditional Analysis* (see, e.g., Ryle 1949; Goodman 1954): An individual x has a disposition D iff x would M when S.

$$D(x) \leftrightarrow (S(x) \rightarrow M(x))$$

However, this analysis has long been recognised as problematic, facing a range of well-known counterexamples: *finks*, *reverse-cycle finks*, *masks* or *antidotes*, and *mimickers* (e.g., Martin 1994; Lewis 1997; Bird 1998). These cases all show that the mere presence of a stimulus is not always sufficient for the manifestation of a disposition or that a disposition may manifest in the absence of the stimulus—there may be interfering or mimicking factors that undermine or simulate the expected outcome. The core issue here is that the antecedent of the conditional must contain more than just the triggering condition; it must encode a rich set of background assumptions or enabling conditions (Choi and Fara 2021; Gozzano 2020).

In our model, as we will discuss in Sect. 4, we will be using a version of the Simple Conditional Analysis as our strategy is to *get more specific*. The approach of getting more specific offers a way out of the most significant objections

[2] The original scenario in Alfieri et al. (2023) uses the phrase "steal some" instead of "grab one". This change is innocuous, since we do not exploit or produce experimental results about the scenario. The term "steal" is perhaps unfortunate, as the act described presents no danger to society or individual gain. In many communities, it is an accepted practice, given also the open gate and the amount of fruits. Still, depending on their moral preferences, individuals act differently, revealing different dispositions.

raised in the literature against the analysis (Manley and Wasserman 2008). Actually, we luckily can assume that all the relevant enabling conditions are explicitly included in the antecedent. This allows us to sidestep many of the standard objections and proceed with a streamlined conditional framework for analyzing dispositions.

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. Let us clarify with an example: the courage of the individual (disposition D) is individuated by the pair of its stimulus condition, which is the dangerous situation (S), and its manifestation, which 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." And, clearly, an individual may be more or less courageous than another. The manifestation of a moral disposition depends on numerous factors, including psychological states and the specific contexts in which the person finds themselves. Even slight variations in these contexts can influence how an individual behaves, potentially preventing a particular disposition from manifesting. Importantly, however, the context-sensitivity of dispositions does not undermine their metaphysical status. Context-sensitivity concerns the conditions under which a disposition manifests, not whether the disposition itself is genuinely possessed. As Ellis (2001) and Bird (2007) argue, dispositions can be essential properties of entities or kinds, even if their manifestation depends on specific background conditions. For example, fragility may only manifest when an object is struck, but that does not make it any less essential to its nature. Similarly, a person may possess courage dispositionally, even if certain contexts inhibit its manifestation. Therefore, the variability of manifestation conditions is fully compatible with essentialist accounts of dispositional properties (Mumford and Anjum 2011a; Gozzano 2020).

Consider a person who is generally known for their kindness. However, in a stressful situation at work, where they feel overwhelmed and unsupported, their usual compassionate nature may not surface. Instead, they may react with frustration or detachment.

There are various theories about dispositions and different versions of those theories, as well as attempts to connect these various versions of dispositionalism with different ethical theories (Anjum et al. 2012; Azzano and Raimondi 2023).

Nevertheless, the standard conception of dispositionalism suffices for our current project. In addition, considering the attempts we mentioned above is not necessary for our purposes, as we are not focused on any particular ethical theory. This minimal version of dispositionalism serves our goal of representing the soft-ethical preferences of individuals.

4 A general model of personal moral values as dispositions

Our primary aim is to elaborate a general model of individuals' soft ethics. Specifically, this model aims to represent the ethical preferences of a single individual and to explore how these preferences guide the individual in selecting and performing the action that best aligns with their values in a given context.

The model adopts a dispositional approach, where actions are manifestations of moral dispositions. These dispositions usually manifest when the individual is exposed to relevant stimuli.

4.1 Moral justifications for actions

We introduce a model of moral justification based on Alfieri et al. (2023). When presented with a scenario (i.e., a soft moral dilemma), an individual brings their own moral dispositions. To justify the action they choose in that scenario, they can explain their decision by stating that the reason for performing that action was:

p_1 based on the fact that the action was consequential on others because of the consequences on others that the actions has;

p_2 consequential on me: because of the consequences on me that the actions has;

p_3 compliant with social norms: because the action complies with norms;

p_4 compliant with personal experience: because the action complies with personal experiences and expertise.

The choice of these four moral principles can be motivated by the fundamental criteria of ethical theorising. The principles can be conveniently placed in a two-dimensional space, see Table 1. One dimension is the axis of personal concerns and social concerns. The other dimension is the

Table 1 Two dimensions and four principles

	consequences	rules
social	p_1	p_3
personal	p_2	p_4

axis of consequences and rules. Using these two meta-values we want to meet the theoretical reflections of both consequentialism and deontology, even if there is no preference on our part for one or the other 'method of ethics' (to call back the title of Henry Sidgwick's *The Methods of Ethics* (Sidgwick 1981)).

The first two principles consider the consequences of the choice adopted by the agent, while the second two consider how the adopted choice conforms or does not conform to either personal rules (principles developed through personal experiences) or to socially accepted conducts.

In order for the subject to express how much the choice they make fits with their own moral dispositions, we allow the subject to justify the action by assigning a value between 1 and 5 to each of the four principles given a specific scenario. What we can deduce is that in the given scenario a certain tuple of values of the four principles is assigned by the subject according to their moral dispositions. By assigning values to these principles, we are treating them as parameters ranging on the set of values 1 to 5. This justifies the choice of calling them "parameters" in the formal model in Sect. 5.

A natural question to ask is whether these four principles are expressive enough. By expressivity, we intend the discriminatory power of the principles when characterising actions. The expressivity of the principles that reflect moral values is, at least partially, in the eyes of the beholder. For, suppose that in the postoffice scenario (Scenario 1) someone declares to be willing to help the clerk and justifies this by assigning the following values: $p_1 \mapsto 5$, $p_2 \mapsto 2$, $p_3 \mapsto 2$, $p_4 \mapsto 5$. So, this person justifies her action by stressing that for her it is important to do what maximises the consequences on others. Presumably, assigning a high value to p_4 is indicative of some sort of gratification she is expecting from helping the clerk.

Clearly, these additional aspects are not explicitly stated neither in the scenarios nor in the principles, but this is a positive aspect. We should keep in mind that we want the principles to facilitate generalisations. In addition, there is a trade-off between expressivity and generalisation: the higher the former, the lower the latter. What is essential then is to find out a proper balance between the capacity of the principles to pinpoint the essential motivational features to perform an action given the context and the individual moral dispositions expressed by the moral justification tuple in other *similar* scenarios.

We may expect, then, that the person whose moral justification tuple in the post office scenario is the one described above, would behave similarly in similar scenarios, that is, in scenarios that can be taken as exemplifying the same type of context (contexts where a single problem, affecting many people, can be resolved by a single action of the subject).

As well as for other kinds of dispositions, moral dispositions are highly context sensitive. Consider the fruit scenario (Scenario 2), where we may assume that if the subject responds YES and decides to take some fruit from the private garden, this may justify them to do similar actions in similar scenarios. Now consider a variation on the original scenario: in the initial case, there are ripe fruits on the trees, no one nearby, and the gate is open. In the variant, ripe fruits are found both on the trees and scattered on the ground everything else being equal. The fallen fruits may suggest that no one is harvesting them. This variation may reinforce the value of the personal experience principle of those who have responded YES to the questionnaire and may move some person from NO to YES, in light of the visible waste of the fruits left to rot.

4.2 Modelling moral dispositions

The general model of moral dispositions is built around a few main elements: world settings, actions and their denotations, and dispositions. Individual agents act within scenarios, which are specific realisations of broader contexts, in accordance with their moral preferences, which are dispositional by nature.

World settings

Modelling world settings is complicated, so we need to model situations that recur and that are similar in some respect. For instance, a broader context is a situation in which some help is needed—all the situations that are similar in this respect are specific realisations of the broader context *needing help*. They satisfy the property 'someone needs help'. There is a similarity relation between all these possible situations.

That is to say, if someone is willing to help, then such a person is one that in a broader context of 'need help', like the post office scenario, is willing to act in support of people hit by the problems.

Actions and their denotations

For moral dispositions to manifest, in addition to a world setting, one needs to consider the set (repertoire) of actions that a person can take. Usually, in the ethical case, we are faced with dilemmas. The set of actions is then composed of two actions (do one action or the other), or simply action that the person must choose to perform or refrain from doing. We can nonetheless consider more articulated scenarios.

What is further needed is a semantics, or denotation, for the actions. As explained in Sect. 4.1, we adopt a set of principles to justify the actions that a person chooses to perform in a given scenario. We also adopt them as the action parameters p_1, ..., p_4 to characterise moral actions. Hence, justifications and action denotations exist in the same space. Justifications can be seen as subjective meanings, while action denotations are the semantics of the action.

Dispositions

We assume that an individual has a set of moral dispositions (e.g., she is generous to some degree; courageous to some degree; etc.). Ideally, these dispositions will be discerned through the subject's preferred course of action (i.e., ethical preferences), and so through the manifestation of some of her dispositions (e.g., helping the staff) in a given world setting that presents some stimuli (e.g., post office with staff experiencing some difficulties).

An action is the manifestation of a disposition, resulting from the interaction between one or more moral dispositions and one or more stimuli (i.e., some properties/disposition of the world setting). The world setting presents the stimuli for the subject's moral dispositions to manifest (e.g., helpfulness) through what the individual considers the most appropriate action (helping). Therefore, an individual acts by selecting one action in a set of other possible actions, by manifesting their moral dispositions.

The action chosen is the one that best represents their dispositions given the scenario in which they are in.

We observed before that human dispositions are highly contextual. Nonetheless, rational decision making also enjoys a certain level of consistency. We assume that people are consistent in similar scenarios. That is, they tend to be helpful in all the settings that exemplify the "need help" context. We know that this assumption could be dissatisfied, but lacking evidence to the contrary, in general people tend to stick with the idea that a helpful person remains a helpful person even if in some scenarios they have acted egoistically. When something like that occurs, we presume that an interfering factor has prevented their natural disposition from occurring. Interferences are present in the dispositional framework and considered as further dispositions, as discussed in Sect. 3—this could be accommodated in a finer-grained model.

Summarising, moral dispositions associate a world setting and a set of action denotations to one of the action denotations. An individual chooses an action with a certain denotation (defined in terms of the four principles), if placed in a given world setting with a set of actions to choose from. Following the conditional analysis of dispositions, we represent the moral disposition and its stimuli in the following way:

$$HasMoralDisposition(x) \leftrightarrow (IsStimulatedForMD(x) \rightarrow ActsAccordingToMD(x))$$

$$IsStimulatedForMD(x) \leftrightarrow (\exists S, X.MakeContextMD(S, X) \wedge InWorldSetting(x, S) \wedge HasActionSet(x, X))$$

That is, an individual has a certain moral disposition iff she acts according to the moral disposition when there is a world setting in which she belongs and a set of actions in her repertoire, that together form a context for the moral disposition. For instance, an individual is helpful iff they would choose the helpful action when confronted to a set of actions in a given world setting that together form a context for manifesting helpfulness.

5 A formal model of ethical dispositions

In this section, we formally define the building blocks of the dispositional model introduced in Sect. 4. This model is used in Sect. 6 for the elicitation and manifestation of human ethical dispositions.

We use standard set-theoretic notation. Let X and Y be two arbitrary sets. The expressions $X \cup Y$ and $X \cap Y$ represent the union and intersection, respectively, of X and Y. The expression X^Y denotes the set of functions from the set Y to the set X. We denote by 2^X the set of subsets of X. We write $|X|$ for the cardinality of X, and $X \subseteq Y$ (resp. $X \subset Y$) to signify that X is a (resp. strict) subset of Y.

We define world settings, actions and action denotations, action denotation maps, action denotation repertoires, and dispositions.

World settings

Let **WSParam** be a set of boolean variables that characterise a *world setting*. What this set is exactly very much depends on the application at hand. Examples of variables can be as diverse as "dangerous situation","chaotic situation", "presence of desirable object", "presence of access control", etc.[3]

A *world setting* SETT is a characterisation of the circumstances of the world and is formalised as a (possibly partial) valuation of the set **WSParam**. That is, SETT : **WSParam** $\longrightarrow \{true, false\}$. **WStt** denotes the set of settings.

Actions and action denotations

Let **Act** be the set of *action types*, such as "help", or "grab", or "not grab". In a given setting, an action is characterised by a set of (weighted) *action parameters*. Let **VParam** be the set of parameters that characterise the moral dimension of an action. Adopting the principles set forth in Alfieri et al. (2023) (and we refer back to Sect. 4.1 for details), the set of parameters characterising actions are:

[3] For more flexibility, one could instead define a world setting to be an extensional database in First-Order or Description Logic Rudolph (2011) over a certain (domain-dependent) signature. However, we do not need this level of sophistication to present our approach.

1. Consequential on others: Describes whether the action has consequences on others or not. Principles p_1.
2. Consequential on me: Describes whether the action has consequences on the actor or it does not. Principle p_2.
3. Compliant with social norms: Describes whether the action follows what could be expected by the norms in the given situation. Principle p_3.
4. Compliant with personal experience: Describes whether the action follows from what we may take as personal experiences and expertise. Principle p_4.

We define $\mathbf{VParam} = \{p_i \mid 1 \leq i \leq 4\}$. In addition, we use a discrete scale ordered from 1 to 5: $\mathbf{Scl} = \{1, 2, 3, 4, 5\}$ to indicate the importance of each parameter.

The set $\mathbf{Scl}^{\mathbf{VParam}}$ is the set of all functions that associate each parameters to an element of **Scl**. An example of one such function is $\{p_1 \mapsto 5, p_2 \mapsto 3, p_3 \mapsto 2, p_4 \mapsto 4\}$ which is depicted in Fig. 1. For convenience, we can also represent the function $\{p_1 \mapsto n_1, p_2 \mapsto n_2, p_3 \mapsto n_3, p_4 \mapsto n_4\}$ as the vector (n_1, n_2, n_3, n_4).

We will think of an element of $\mathbf{Scl}^{\mathbf{VParam}}$ as of an *action denotation*.

Action denotation maps

An action may be characterised by different functions from **VParam** to **Scl** when situated in different world settings.

An *action denotation map* ADM is a function that associates a action denotation (a function from **VParam** to **Scl**) to a pair made up of a world setting and an action:

$$\text{ADM} : \mathbf{WStt} \times \mathbf{Act} \longrightarrow \mathbf{Scl}^{\mathbf{VParam}} \ .$$

For simplicity of presentation, we assume that this map is objective. However, in practice, stakeholders can resort to individual maps tailored to their needs.

Example 1 Suppose that SETT(postoffice) describes the world setting of Scenario 1 ("public service place", "readiness to leave", "broken queue management system", "chaotic situation"), and α is the action of stopping and helping the clerk. ADM(SETT(postoffice), α) could be $\{p_1 \mapsto 5, p_2 \mapsto 2, p_3 \mapsto 4, p_4 \mapsto 1\}$ (also noted (5, 2, 4, 1)),

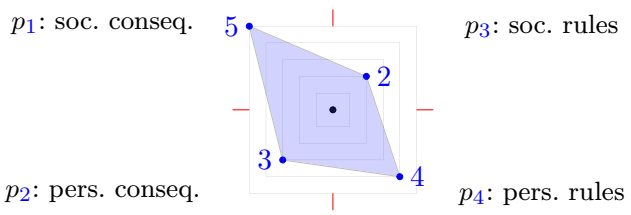

Fig. 1 Geometric representation of the action denotation $\{p_1 \mapsto 5, p_2 \mapsto 3, p_3 \mapsto 2, p_4 \mapsto 4\}$

thus indicating that the surveyor intends that helping the clerk in this setting denotes an action that has primarily social consequences and is based on social principles.

Suppose that SETT(fruit) describes the world setting of Scenario 2 ("private property", "no one observing", "presence of access control", "access control not activated", "presence of desirable object"), and α is the action of not going in (and not grabbing a fruit). ADM(SETT(fruit), α) could be $\{p_1 \mapsto 5, p_2 \mapsto 2, p_3 \mapsto 5, p_4 \mapsto 4\}$. Let now α' be the action of entering and grabbing a fruit. ADM(SETT(fruit), α') could be $\{p_1 \mapsto 1, p_2 \mapsto 4, p_3 \mapsto 1, p_4 \mapsto 3\}$.

As a variant of the previous world setting, suppose that it is not the case that "private property" and instead "public property". ADM(SETT, α') could then be $\{p_1 \mapsto 1, p_2 \mapsto 4, p_3 \mapsto 3, p_4 \mapsto 3\}$, as the same concrete action would not conflict as much with the social norms as in the world setting of the scenario.

Action repertoires

An *action repertoire* is simply a subset of **Act**, that is, an element of $2^{\mathbf{Act}}$. An *action denotation repertoire* is a set of action denotations, that is, an element of $2^{(\mathbf{Scl}^{\mathbf{VParam}})}$.

We will assume that every action in an action repertoire resolves to a *distinct* action denotation. That is, given a world setting SETT, the function ADM(SETT, x) is a bijection between the action repertoire **rep** and the set $\{\text{ADM}(\text{SETT}, \alpha) \mid \alpha \in \mathbf{rep}\}$.

Example 2 The set {"grab", "not grab"} is an action repertoire. The set $\{(1, 4, 1, 3), (5, 2, 5, 4)\}$ is an action denotation repertoire. Informally, this could be the action denotation repertoire of the actions "grab" and "not grab" in the world setting of Scenario 2.

Dispositions

As anticipated in Sects. 3 and 4.2, we take advantage of the conditional analysis of dispositions.

A *disposition* is formalised as a mapping of a pair consisting of

- a world setting SETT, and
- an action denotation repertoire χ

to an element X of χ:

$$\text{SETT}, \chi \mapsto X \ .$$

This formalises the fact that in a *particular* world setting SETT, and with the possibility to choose among a *particular* set of action denotations χ, the individual has a disposition to choose a certain action denotation $X \in \chi$.

We note that choosing one item from a set of items is a typical topic of study in decision theory (Luce and Raiffa

1957), and AI. Further investigations to put our model into practice could take advantage of existing results and techniques. However, many results from decision theory are based on rationality axioms. One must be careful in applying assumptions too early as they may discard some phenomena otherwise observed in human action. Some rationality assumptions can be made, for example, by applying a principle that is often called *independence of irrelevant alternative*. If an individual chooses a over b in X, and $a, b \in Y \subset X$, then they choose also a over b in Y. But this does not account for 'menu effects': one may have a tendency to choose the apple if presented with an apple, a peach, and a pear, but have a tendency to choose the peach if presented with an apple and a peach, and no pear. Menu effects have been particularly observed in self-control-based decision making, as in Borah and Garg (2023).

Henceforth, we make no claims about the logical nature of the $\mapsto$ connective. The reader should in particular refrain from interpreting it as the material implication of classical logic. For instance, it is possible that an individual has the disposition $\text{SETT}, \{X, Y\} \mapsto X$, but also the disposition $\text{SETT} \cup \{w\}, \{X, Y\} \mapsto Y$ or $\text{SETT}, \{X, Y, Z\} \mapsto Y$.

6 Disposition elicitation and manifestation

We explain how the dispositions can be elicited and how they can manifest.

6.1 Elicitation tool

As we anticipated, questionnaires are one of the main tools for gathering user experiences, feedback, and preferences. They help gather structured data in a standardised way, facilitating the analysis and use of answers.

A *questionnaire* contains a set of scenarios. For each scenario, one expects an *answer* from a user.

Scenarios

A *scenario* s, designed by a surveyor, has a 'public' part and an 'implementation' part. The public part is like in Alfieri et al. (2023): a description of the world setting in plain English, a description of the problem in plain English, and a repertoire of actions $\mathbf{rep}(s)$ (possibly a dichotomous choice between doing an action and refraining from doing the action).

The implementation part specifies:

- a world setting $\text{SETT}(s)$ as an element of **WStt**. This should be a symbolic representation of the description of the setting of the public part, in the language of the parameters in **WSParam**.
- for every action α in the action repertoire $\mathbf{rep}(s)$, an action denotation $\text{ADM}(\text{SETT}, \alpha)$. This should be the denotation of the concrete actions described in the public part, that is, a function in $\mathbf{Scl}^{\mathbf{VParam}}$.

Example 3 The public part of the scenario `fruits` is simply the English description in Scenario 2.

The implementation part consists of the world setting $\text{SETT}(\texttt{fruit})$ and an action denotation for each action, "not grab", and "grab", all supposedly formalised by the designers of the scenario.

The world setting $\text{SETT}(\texttt{fruit})$ and the actions are as formalised in Example 1, second paragraph. The world setting $\text{SETT}(\texttt{fruit})$ consists of the following elements of **WSParam**, "private property", "no one observing", "presence of access control", "access control not activated", "presence of desirable object". The action denotations are $\text{ADM}(\text{SETT}(\texttt{fruit}), \text{"grab"}) = (1, 4, 1, 3)$, and $\text{ADM}(\text{SETT}(\texttt{fruit}), \text{"not grab"}) = (5, 2, 5, 4)$.

Answers

An *answer* to a scenario s (with world setting $\text{SETT}(s)$) by an individual is made of:

- an action α in $\mathbf{rep}(s)$,
- a justification $\text{JUST} \in \mathbf{Scl}^{\mathbf{VParam}}$.

An answer (α, JUST) is *consistent* when the justification JUST for action α 'bears some similarity' with the action denotation $\text{ADM}(\text{SETT}(s), \alpha)$ specified by the surveyor. When that is the case, we write $\mathit{consistent}(\text{SETT}(s), \alpha, \text{JUST})$.

A candidate measure for this similarity is the *Jaccard–Ruzicka index*, defined for two arbitrary n-vectors of positive numbers $U = (u_i)_{1 \le i \le n}$ and $V = (v_i)_{1 \le i \le n}$ as $\mathit{Ruz}(U, V) = 1$ when $U = V = (0, \dots, 0)$, and $\mathit{Ruz}(U, V) = \frac{\sum_i \min(u_i, v_i)}{\sum_i \max(u_i, v_i)}$ otherwise. It ranges between 0 and 1, where 0 represents the (limit) maximum dissimilarity, and 1 represents equality.

We can thus define that $\mathit{consistent}(\text{SETT}(s), \alpha, \text{JUST})$ is true iff

$$\mathit{Ruz}(\text{JUST}, \text{ADM}(\text{SETT}(s), \alpha)) > 1 - \epsilon \ ,$$

for some $0 < \epsilon \le 1$.

6.2 Ethical disposition elicitation

We have a scenario s and an answer (action α in $\mathbf{rep}(s)$ and justification).

If the answer is not consistent, no disposition is elicited. If the answer is consistent, we elicit the disposition by mapping

$$(\text{SETT}(s), \{\text{ADM}(\text{SETT}(s), \beta) \mid \beta \in \mathbf{rep}(s)\}) \ ,$$

with

$\text{ADM}(\text{SETT}(s), \alpha)$.

That is, in a scenario satisfying the world setting $\text{SETT}(s)$, with the set $\{\text{ADM}(\text{SETT}(s), \beta) \mid \beta \in \mathbf{rep}(s)\}$ of social action denotations to choose from, the individual has a *disposition* to choose $\text{ADM}(\text{SETT}(s), \alpha)$. This is illustrated in Fig. 2.

Example 4 Figure 2 illustrates a case of moral disposition elicitation in Scenario 2.

The world setting $\text{SETT}(\mathsf{fruit})$ is as described in the previous examples. The action denotations are $\text{ADM}(\text{SETT}(\mathsf{fruit}), \text{“grab”}) = (1, 4, 1, 3)$, and $\text{ADM}(\text{SETT}(\mathsf{fruit}), \text{“not grab”}) = (5, 2, 5, 4)$.

Suppose that the user chooses $\alpha =$ “grab”, with justification $\text{JUST} = \{p_1 \mapsto 1, p_2 \mapsto 5, p_3 \mapsto 1, p_4 \mapsto 5\}$.

Suppose also that $\epsilon = 0.3$. We have $Ruz(\text{JUST}, \text{ADM}(\text{SETT}(\mathsf{fruit}), \alpha)) = \frac{1+4+1+3}{1+5+1+5} = 0.75 > 1 - \epsilon$. So $consistent(\text{SETT}(\mathsf{fruit}), \alpha, \text{JUST})$ is true, indicating that the answer is consistent. Thus, the following disposition is elicited:

$$\begin{aligned}(\text{SETT}(\mathsf{fruit}), \{&\{p_1 \mapsto 1, p_2 \mapsto 4, p_3 \mapsto 1, p_4 \mapsto 3\},\\ &\{p_1 \mapsto 5, p_2 \mapsto 2, p_3 \mapsto 5, p_4 \mapsto 4\}\})\\ \mapsto\ &\{p_1 \mapsto 1, p_2 \mapsto 4, p_3 \mapsto 1, p_4 \mapsto 3\}\ .\end{aligned}$$

One can highlight that elicited dispositions pertain to action denotations, not concrete actions. This level of abstraction is useful for generalisation and disposition manifestation in situation not experienced before.

6.3 Disposition manifestation

Suppose one has elicited a set of dispositions Δ for an agent. Now, the agent is in a world setting SETT and confronted to a set of actions $\mathbf{rep} \subseteq \mathbf{Act}$. The latter corresponds to the set of action denotations $\mathbf{rep}_{\text{ADM}} = \{\text{ADM}(\text{SETT}(s), \beta) \mid \beta \in \mathbf{rep}\}$.

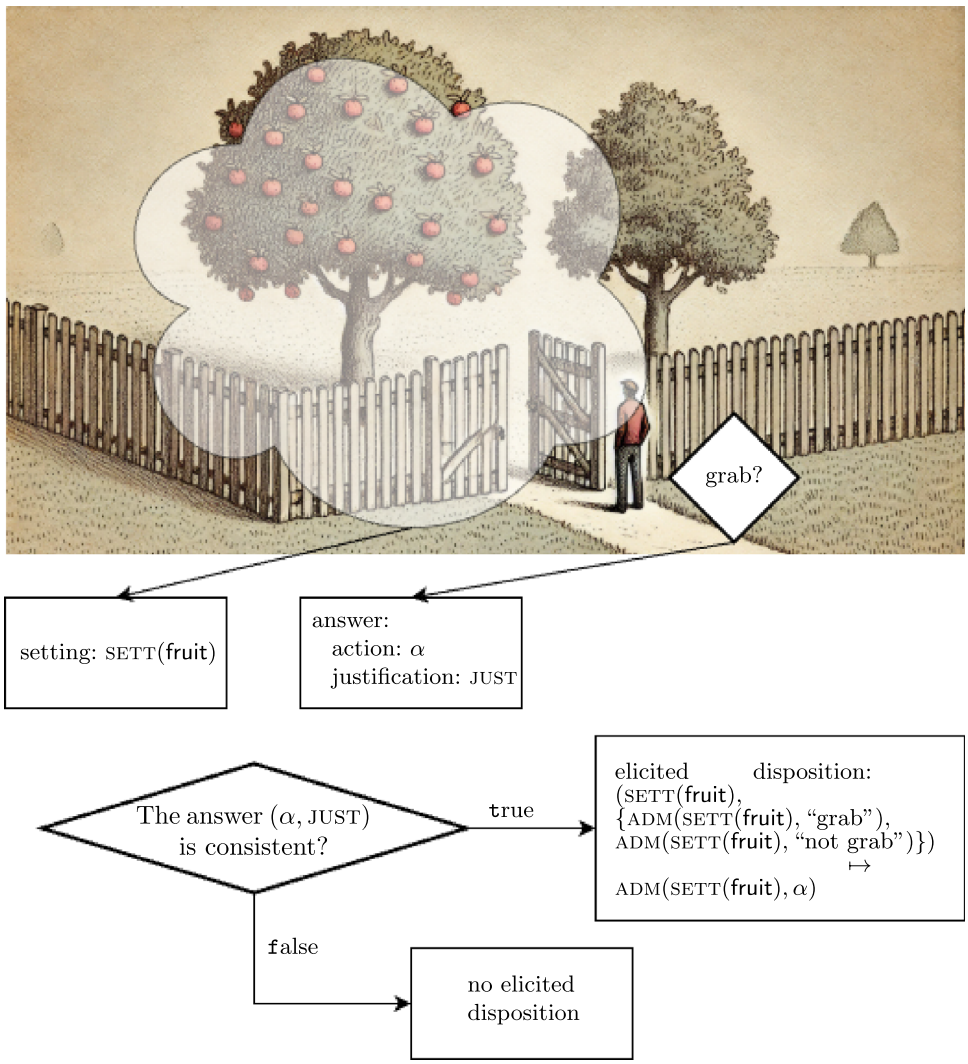

Fig. 2 From scenarios to moral disposition. Depiction of Scenario 2. Some details are presented Example 4

A Δ-*disposition manifestation mechanism* is a function of the set of elicited dispositions Δ, and such that for every pair consisting of

- SETT, and
- $\mathbf{rep}_{\text{ADM}}$

an element X of $\mathbf{rep}_{\text{ADM}}$ is returned. That is,

$$\Delta, (\text{SETT}, \mathbf{rep}_{\text{ADM}}) \mapsto X \ .$$

If the pair $(\text{SETT}, \mathbf{rep}_{\text{ADM}})$ appears as such and exactly in one elicited disposition of Δ, then it should manifest, and the corresponding action should be chosen. However, in general, the pair $(\text{SETT}, \mathbf{rep}_{\text{ADM}})$ might not appear exactly in an elicited disposition. Whether a disposition should manifest, and if yes, which one, is dependent on the application at hand.

Stakeholders (recommendation system designers, users configuring their digital assistant, etc.) may decide what best decision-making approach they reckon should apply. In the interest of generality, we do not commit to one in particular. We remark that the present model may well accommodate any disposition manifestation mechanism, and we now discuss a few possibilities. If no elicited disposition exists, where $(\text{SETT}, \mathbf{rep}_{\text{ADM}})$ exactly occurs, the system could: do nothing, or choose action in $\mathbf{rep}$ uniformly at random. Or the system could adopt a decision-making procedure that obeys some "rationality axioms". To illustrate this, we list a few candidates for such axioms, staying faithful to the intended generality of the model.

1. If Δ contains some dispositions $s, \chi \mapsto X$, where $\chi \subseteq \mathbf{rep}_{\text{ADM}}$, and choose an action 'recommended' by one of them which is then manifested.
2. If Δ contains some dispositions $s, \chi \mapsto X$, where $\mathbf{rep}_{\text{ADM}} \subseteq \chi$, and $X \in \mathbf{rep}_{\text{ADM}}$, choose an action 'recommended' by them, which is then manifested.
3. When more than one disposition in Δ could manifest, choose the action of the one with a world setting which is 'closer' to SETT. Since our settings are simple valuations, the Hamming distance could be used. In more complex world setting characterisations (e.g., with Description Logic), more refined methods like structure-mapping engines (Falkenhainer et al. 1989) can also be exploited.
4. When more than one disposition in Δ could manifest, choose the action of the one with an action denotation repertoire which is 'closer' to $\mathbf{rep}_{\text{ADM}}$ (e.g., some elaboration on the Jaccard–Ruzicka index).

We believe that an axiomatic characterisation of moral disposition manifestation, following the tradition of decision theory (Luce and Raiffa 1957), would be fruitful to guide the design of personal ethical decision systems and their evaluation.

Example 5 Figure 3 illustrates the manifestation of dispositions. The scenario is one, where an ethically aware digital personal assistant is interacting with an autonomous queue management system on behalf of its user.

The concrete actions are "swap" (swap position with person in wheelchair), and "noswap" (do nothing). Indicatively, we assume that in this world setting SETT, "swap" can be denoted as a social action having consequence on others, and on the actor, and is compliant with norms and personal experience. The concrete action "noswap" can instead be denoted as a social action

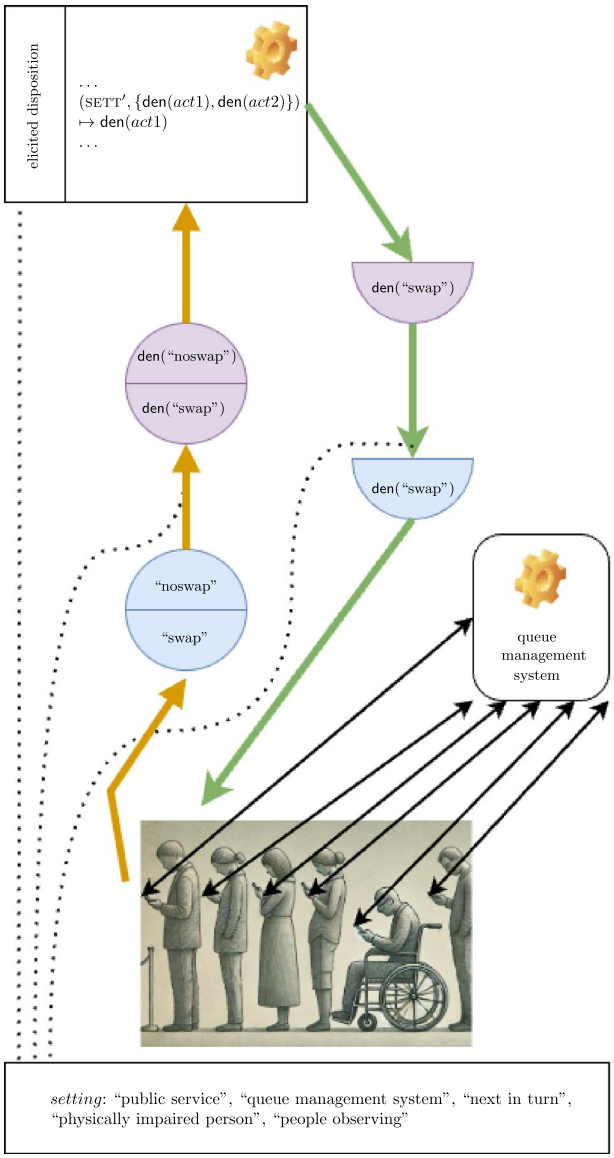

Fig. 3 From moral dispositions to action. Some details are presented in Example 5

that has some consequence on others, and on the actor. That is, it could be, $\mathsf{den}(\text{“swap”}) = \text{ADM}(\text{SETT}, \text{“swap”}) = \{p_1 \mapsto 4, p_2 \mapsto 4, p_3 \mapsto 4, p_4 \mapsto 5\}$, $\mathsf{den}(\text{“noswap”}) = \text{ADM}(\text{SETT}, \text{“noswap”}) = \{p_1 \mapsto 4, p_2 \mapsto 4, p_3 \mapsto 2, p_4 \mapsto 2\}$.

The figure illustrates a case, where the two world settings SETT and SETT′ ‘bear some similarity’, and so do $\mathsf{den}(\text{“swap”})$ and $\mathsf{den}(act1)$, and $\mathsf{den}(\text{“noswap”})$ and $\mathsf{den}(act2)$. Having the elicited disposition $(\text{SETT}', \{\mathsf{den}(act), \mathsf{den}(act2)\}) \mapsto \mathsf{den}(act1)$, the mechanism yields the concrete action “swap”.

7 Outlook

We elaborated a general model of personal moral values. It can serve at gathering information about personal soft ethics of human agents from a questionnaire, elicit ethical preferences as dispositions, and use these dispositions to individuate preferred behaviours, i.e., manifestations.

The model of personal moral values described in Sects. 4, 5 and 6 is readily amenable to both specialisations and generalisations. Regarding specialisation, the function that evaluates the consistency of actions and justifications during disposition elicitation in Sect. 6.1, can be adjusted. At one extreme, it can be trivialised to always return true, thereby eliciting all answers into a disposition. At the other extreme, it can be strengthened to elicit only perfectly justified answers. In fact, under the proposed definition, this corresponds to setting ϵ to 1 and 0, respectively. In addition, the procedure described in Sect. 6.3 for disposition manifestation can be instantiated as a concrete mechanism for a specific application domain. Regarding generalisation, the model allows for abstraction from both the set of principles and the scale of their valuations. Although we used the four principles introduced in Sect. 4.1 and a scale from 1 to 5 to provide a semantics for actions, the sets denoted **VParam** and **Scl** in Sect. 5 can be redefined as needed without altering the fundamental idea of the model.

The limits of empowerment

We recognise that user empowerment, although frequently seen as a design ideal, is not ethically neutral. As Gabriel (2020) points out, human desires, beliefs, and preferences can be biased, harmful, or in conflict with shared societal norms. For this reason, value alignment goes beyond simply recording what users want: it must also involve a critical examination of the moral legitimacy of those wants. In our work, we do not treat empowerment as an unquestioned good or as an end in itself. It is not about creating superhumans; rather, it is about staying human, complete with all our imperfections, within the space of soft ethics. We aim not for empowerment tout-court, but within the interaction with digital technologies, to rebalance power structures that too often leave the user voiceless. This process necessarily includes the possibility of encountering ethically borderline or socially undesirable preferences. Indeed, we are aware that some preferences may perhaps conflict with legal norms or established ethical standards. However, eliciting such preferences is not the same as endorsing them, let alone implementing them. On the contrary, we believe that systems will include hard ethical constraints: normative boundaries that are non-negotiable and built-in. These constraints will prevent the manifestation of preferences that in some contexts may become unlawful, in line with hard ethics. Still, even the morally dubious preferences can be valuable. They provide insight into the complexity of individual moral landscapes and help build richer, more realistic moral profiles. In this way, preference elicitation is not a vehicle for unchecked personalisation. Rather, it becomes a tool for critical reflection and an opportunity to explore the tension between individual morality and collective norms, and ultimately contributes to a more ethically grounded system design.

Future work

The first avenue for future work involves developing a formal language and reasoning services to represent the dispositions elicited and manifested, as described in this paper. This is crucial for constructing a computational model of individual soft ethics as dispositions and for adapting our proposal to real-world systems. We believe that classical logic or logic programming will be particularly fruitful in this regard. In Troquard et al. (2024), some progress has already been made by capturing social, legal, ethical, empathetic, and cultural rules from Townsend et al. (2022). A key feature of SLEEC rules is their ability to account for exceptions to normative rules: “You should do *A* unless *C* is the case, in which case you should do *B*”. This reflects a non-monotonic behaviour, which is also characteristic of dispositions. Despite their defeasible nature, these rules are simple enough to be efficiently captured using classical logic. However, classical logic may be too coarse to adequately capture finer-grained dispositions. Therefore, we will also explore the use of probabilistic rules or fuzzy logic (Zadeh 1988).

We already mentioned in Sects. 5 and 6.3 the possibility of an axiomatic analysis of the decision processes involved in the manifestation of moral disposition. This is also intimately connected to the perspective of developing a formal logic for moral dispositions, as discussed before. We believe this to be a fascinating topic that has been little explored. This would further inform a formal and computational treatment of our model.

Ultimately, we want to use the gathered moral dispositions to create a software profile that enhances human abilities by respecting their ethical choices when they interact with autonomous systems. To this aim, we plan to

investigate the use of inductive logic programming (ILP) (Cropper and Dumancic 2022). Every SETT, $\chi \mapsto X$ in a set of elicited dispositions Δ becomes an example, and a Δ-disposition manifestation mechanism (c.f., Sect. 6.3) can thus be learnt automatically from them.

Author contributions All authors whose names appear on the submission made substantial contributions to the conception of the work, drafted the work and revised it critically for important intellectual content approved the version to be published. All authors agree to be accountable for all aspects of the work.

Funding Open access funding provided by Università degli Studi dell'Aquila within the CRUI-CARE Agreement. The authors acknowledge the support of the MUR (Italy) Department of Excellence 2023–2027 for GSSI, and of the PRIN project 2022JKA4SL—HALO: etHical-aware AdjustabLe autOnomous systems.

Data availability No data sets were generated or analysed during the current study.

Code availability Not applicable.

Materials availability Not applicable.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval Our research focuses on the design of autonomous systems that integrate decision-making with ethical considerations. The future questionnaires and systems designed based on the ideas from this research will be required to adhere to strict ethical guidelines. The present research itself does not raise any ethical concerns.

Consent for publication All authors agreed with the content and all gave explicit consent to submit.

References

Alfieri C, Donati D, Gozzano S et al (2023) Ethical preferences in the digital world: the EXOSOUL questionnaire. In: Lukowicz P, Mayer S, Koch J et al (eds) HHAI 2023: augmenting human intellect—proceedings of the second international conference on hybrid human-artificial intelligence, June 26–30, 2023, Munich, Germany. Vol 368. Frontiers in Artificial Intelligence and Applications. IOS Press, pp 290–299. https://doi.org/10.3233/FAIA230092

Anjum RL, Lie SAN, Mumford S (2012) Chap: Dispositions and ethics. In: Powers and capacities in philosophy. Routledge, London

Atari M, Haidt J, Graham J et al (2023) Morality beyond the weird: how the nomological network of morality varies across cultures. J Pers Soc Psychol 125(5):1157–1188. https://doi.org/10.1037/pspp0000470

Autili M, Ruscio DD, Inverardi P et al (2019) A software exoskeleton to protect and support citizen's ethics and privacy in the digital world. IEEE Access 7:62011–62021. https://doi.org/10.1109/ACCESS.2019.2916203

Autili M, De Sanctis M, Inverardi P et al (2025) Engineering digital systems for humanity: a research roadmap. ACM Trans Softw Eng Methodol. https://doi.org/10.1145/3712006. (**accepted**)

Awad E, Dsouza S, Kim R et al (2018) The moral machine experiment. Nature 563:59–64

Azzano L, Raimondi A (2023) Vices, virtues, and dispositions. TheoLogica Int J Philos Relig Philos Theol. https://doi.org/10.14428/thl.v7i2.67873

Bird A (1998) Dispositions and antidotes. Philos Q 48(192):227–234

Bird A (2007) Nature's metaphysics: laws and properties. Oxford University Press, Oxford

Boltz N, Yaman SG, Inverardi P et al (2024) Human empowerment in self-adaptive socio-technical systems. In: Baresi L, Ma X, Pasquale L (eds) Proceedings of the 19th international symposium on software engineering for adaptive and self-managing systems, SEAMS 2024, Lisbon, Portugal, April 15–16, 2024. ACM, pp 200–206

Borah A, Garg R (2023) Reference-dependent self-control: menu effects and behavioral choices. J Econ Behav Org 211:129–145. https://doi.org/10.1016/j.jebo.2023.04.027

Choi S, Fara M (2021) Dispositions. In: Zalta EN (ed) The stanford encyclopedia of philosophy (Spring 2021 Edition). https://plato.stanford.edu/archives/spr2021/entries/dispositions/

Cropper A, Dumancic S (2022) Inductive logic programming at 30: a new introduction. J Artif Intell Res 74:765–850. https://doi.org/10.1613/JAIR.1.13507

Curry OS, Jones Chesters M, Van Lissa CJ (2019) Mapping morality with a compass: testing the theory of 'morality-as-cooperation' with a new questionnaire. J Res Pers 78:106–124. https://doi.org/10.1016/j.jrp.2018.10.008

Driver J (2022) Moral theory. In: Zalta EN, Nodelman U (eds) The Stanford encyclopedia of philosophy, Fall, 2022nd edn. Metaphysics Research Lab, Stanford University

Ellis B (2001) Scientific essentialism. Cambridge University Press, Cambridge

European Commission and Directorate-General for Research and Innovation and European Group on Ethics in Science and New Technologies (2018) Statement on artificial intelligence, robotics and "autonomous" systems —Brussels, 9 March 2018. Publications Office. https://doi.org/10.2777/531856

European Data Protection Supervisor (2015) Towards a new digital ethics: data, dignity and technology. https://www.edps.europa.eu/sites/default/files/publication/15-09-11_data_ethics_en.pdf

European Parliament (2024) Artificial Intelligence Act. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

Falkenhainer B, Forbus KD, Gentner D (1989) The structure-mapping engine: algorithm and examples. Artif Intell 41(1):1–63

Feffer M, Skirpan M, Lipton ZC et al (2023) From preference elicitation to participatory ML: a critical survey & guidelines for future research. In: Rossi F, Das S, Davis J et al (eds) Proceedings of the 2023 AAAI/ACM conference on AI, ethics, and society, AIES 2023, Montréal, QC, Canada, August 8–10, 2023. ACM, pp 38–48

Feng N, Marsso L, Yaman SG et al (2024) Analyzing and debugging normative requirements via satisfiability checking. In: International conference on software engineering. ACM, pp 214:1–214:12. https://doi.org/10.1145/3597503.3639093

Floridi L (2018) Soft ethics, the governance of the digital and the general data protection regulation. Philos Trans R Soc A 376(2133):20180081. https://doi.org/10.1098/rsta.2018.0081
Floridi L, Taddeo M (2016) What is data ethics? Philols Trans R Soc A 374(2083):20160360. https://doi.org/10.1098/rsta.2016.0360
Friedman B, Hendry DG (2019) Value sensitive design: shaping technology with moral imagination. MIT Press, Cambridge
Friedman B, Hendry DG, Borning A (2017) A survey of value sensitive design methods. Found Trends Hum Comput Interact 11(2):63–125. https://doi.org/10.1561/1100000015
Fukuyama F, Richman B, Goel A (2021) How to save democracy from technology: ending big tech's information monopoly. Foreign Aff 100(2):98–110
Gabriel I (2020) Artificial intelligence, values, and alignment. Mind Mach 30(3):411–437. https://doi.org/10.1007/s11023-020-09539-2
Goodman N (1954) Fact, fiction, and forecast. Harvard University Press, Cambridge
Gozzano S (1997) Intenzionalità, contenuto e comportamento. Armando Editore, Rome
Gozzano S (2020) Necessitarianism and dispositions. Metaphysica 21(1):1–23
Graham J, Nosek BA, Haidt J et al (2011) Mapping the moral domain. J Pers Soc Psychol 101(2):366–385. https://doi.org/10.1037/a0021847
Graham J, Haidt J, Koleva S et al (2013) Chapter two—Moral foundations theory: the pragmatic validity of moral pluralism. In: Devine P, Plant A (eds) Advances in experimental social psychology, vol 47. Academic Press, Cambridge, pp 55–130. https://doi.org/10.1016/B978-0-12-407236-7.00002-4
Inverardi P (2019) The European perspective on responsible computing. Commun ACM 62(4):64. https://doi.org/10.1145/3311783
Lewis D (1997) Finkish dispositions. Philos Q 47:143–158. https://doi.org/10.1111/1467-9213.00052
Luce RD, Raiffa H (1957) Games and decisions. Wiley, New York
Manley D, Wasserman R (2008) On linking dispositions and conditionals. Mind 117(465):59–84
Martin CB (1994) Dispositions and conditionals. Philos Q 44(174):1–8. https://doi.org/10.2307/2220143
Mougouei D, Perera H, Hussain W et al (2018) Operationalizing human values in software: a research roadmap. In: Leavens GT, Garcia A, Pasareanu CS (eds) Proceedings of the 2018 ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering, ESEC/SIGSOFT FSE 2018, Lake Buena Vista, FL, USA, November 04–09, 2018. ACM, pp 780–784. https://doi.org/10.1145/3236024.3264843
Mumford S, Anjum RL (2011a) Getting causes from powers. Oxford University Press, Oxford
Mumford S, Anjum RL (2011b) Chap: Dispositional modality. Lebenswelt und Wissenschaft. De Gruyter, Berlin, pp 380–394
Nyholm S (2023) Responsibility gaps, value alignment, and meaningful human control over artificial intelligence. In: Risk and responsibility in context, 1st edn. Routledge. p 23. Open Access, CC BY-NC. Funder: FWF Austrian Science Fund. https://doi.org/10.4324/9781003276029-3
Peterson M (2009) An introduction to decision theory. Cambridge University Press, Cambridge
Rudolph S (2011) Foundations of description logics. Springer, Berlin Heidelberg, pp 76–136. https://doi.org/10.1007/978-3-642-23032-5_2
Ryle G (1949) The concept of mind. Penguin, London
Schwartz SH (1992) Chapter 25: Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries. In: Zanna MP (ed) Advances in experimental social psychology, vol 25. Academic Press, Cambridge, pp 1–65. https://doi.org/10.1016/S0065-2601(08)60281-6
Schwartz SH (2012) An overview of the Schwartz theory of basic values. Online Read Psychol Cult 2(1):1–20
Schwartz SH, Cieciuch J, Vecchione M et al (2012) Refining the theory of basic individual values. J Pers Soc Psychol 103(4):663–688
Shahin M, Hussain W, Nurwidyantoro A et al (2022) Operationalizing human values in software engineering: a survey. IEEE Access 10:75269–75295. https://doi.org/10.1109/ACCESS.2022.3190975
Sidgwick H (1981) Methods of ethics, 7th edn. Hackett Publishing Co, Indianapolis
Tolmeijer S, Kneer M, Sarasua C et al (2021) Implementations in machine ethics: a survey. ACM Comput Surv. https://doi.org/10.1145/3419633
Townsend BA, Paterson C, Arvind TT et al (2022) From pluralistic normative principles to autonomous-agent rules. Mind Mach 32(4):683–715. https://doi.org/10.1007/S11023-022-09614-W
Townsend B, Parnell KJ, Yaman SG et al (2025) Normative conflict resolution through human–autonomous agent interaction. J Respons Technol 21:100114. https://doi.org/10.1016/j.jrt.2025.100114 (https://www.sciencedirect.com/science/article/pii/S2666659625000101)
Troquard N, De Sanctis M, Inverardi P, et al (2024) Social, legal, ethical, empathetic, and cultural rules: compilation and reasoning. In: Wooldridge MJ, Dy JG, Natarajan S (eds) 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 2014, February 20–27, 2024, Vancouver, Canada. AAAI Press, pp 22385–22392. https://doi.org/10.1609/AAAI.V38I20.30245
Vallor S (2016) An introduction to data ethics. Markkula Center for Applied Ethics, Santa Clara
Vetter B (2015) Potentiality: from dispositions to modality. Oxford University Press, Oxford, New York
Wallach W, Allen C (2009) Moral machines: teaching robots right from wrong. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780195374049.001.0001
Yaman SG, Ribeiro P, Cavalcanti A et al (2025) Specification, validation and verification of social, legal, ethical, empathetic and cultural requirements for autonomous agents. J Syst Softw 220:112229. https://doi.org/10.1016/J.JSS.2024.112229
Zadeh LA (1988) A computational theory of dispositions. In: Turksen IB, Asai K, Ulusoy G (eds) Comput Integr Manuf. Springer, Berlin, Heidelberg, pp 215–241