Moral agents as relational systems: The Contract-Based Model of Moral Cognition for AI

Luis Marcos[a], **Serena Marchesi**[b], **Agnieszka wykowska**[b] **and Clara Pretus**[c;*]
[a]Department de Psiquiatria i Medicina Legal, Universitat Autònoma de Barcelona
[b]Social Cognition in Human-Robot Interaction, Italian Institute of Technology
[c]Fundació Institut Hospital del Mar d'Investigacions Mèdiques, Barcelona, Spain
ORCiD ID: Serena Marchesi https://orcid.org/ORCID:0000-0001-9931-156X,
Agnieszka wykowska https://orcid.org/ORCID:0000-0003-3323-7357,
Clara Pretus https://orcid.org/ORCID:0000-0003-2172-1184

Abstract. As artificial systems are becoming more prevalent in our daily lives, we should ensure that they make decisions that are aligned with human values. Utilitarian algorithms, which aim to maximize benefits and minimize harm fall short when it comes to human autonomy and fairness since it is insensitive to other-centered human preferences or how the burdens and benefits are distributed, as long as the majority benefits. We propose a Contract-Based model of moral cognition that regards artificial systems as relational systems that are subject to a social contract. To articulate this social contract, we draw from contractualism, an impartial ethical framework that evaluates the appropriateness of behaviors based on whether they can be justified to others. In its current form, the Contract-based model characterizes artificial systems as moral agents bound to obligations towards humans. Specifically, this model allows artificial systems to make moral evaluations by estimating the relevance each affected individual assigns to the norms transgressed by an action. It can also learn from human feedback, which is used to generate new norms and update the relevance of different norms in different social groups and types of relationships. The model's ability to justify their choices to humans, together with the central role of human feedback in moral evaluation and learning, makes this model suitable for supporting human autonomy and fairness in human-to-robot interactions. As human relationships with artificial agents evolve, the Contract-Based model could also incorporate new terms in the social contract between humans and machines, including terms that confer artificial agents a status as moral patients.

1 The role of contractualism in artificial systems

After the death of a teenage girl in 2017, a compilation of "depression pins you might like" and tips on how to hide mental illness from family members was discovered in her email. Recently, a British court concluded that the teenager died in part due to the detrimental impact of online content, marking a potential first instance in which internet corporations have been held legally responsible for a suicide (New York Times, 2022). These developments take place amidst increasing public awareness that social media algorithms promote eating disorders, self-harm, and suicide among teens (AP, 2022). As artificial systems are becoming more prevalent in our daily lives, we should ensure that they make decisions that are aligned with human values, especially when these decisions have a significant impact on individuals and society as a whole. But how can we build artificial systems that care for humans?

The ethical principles that guide the design and implementation of artificial systems can vary widely depending on the specific application and context of the system. A common approach AI developers adopt is consequentialism, a normative ethical theory that evaluates the morality of actions based on their outcomes or consequences (Moor, 1999). For example, artificial systems used in healthcare may be designed to prioritize patient well-being and avoid harm, while those used in finance may be designed to optimize profit and minimize risk. Consequentialist algorithms can aggregate benefits across a pool of individuals and find the parameters that maximize benefits for a majority of individuals [2]. Thus, because of its capacity to handle large numbers of people, consequentialism is particularly helpful for organizations.

Algorithms based on utilitarianism, a form of consequentialism that aims to maximize human pleasure or happiness, can successfully navigate two pillars of applied ethics: beneficence and non-maleficence. However, they fall short when it comes to human autonomy and fairness, the two other applied ethics principles. Because utilitarianism is insensitive to human preferences beyond well-being and suffering, it is limited in its understanding of human autonomy, or the ability of individuals to make choices for themselves. It also disregards the principle of fairness, or how the burdens and benefits are distributed, as long as the majority benefits (but see also [9]). Relatedly, utilitarianism doesn't consider concepts such as intentions or accountability. Because these are aspects that are central to human morality, artificial systems based on consequentialism may feel dehumanizing to humans [14]. As a result, people may perceive reduced moral agency and low trust in AI.

So how can we build systems that are centered on human autonomy and fairness, and feel less dehumanizing? In the current paper, we present a model of moral cognition for AI based on contractualism. Contractualism is a recent normative ethical theory that attempts to derive moral content from an agreement between free and equal individuals (Scanlon, 2008). Beyond Rousseau's idea of a social contract between a state and its citizens, contractualism focuses on individuals' rights and obligations negotiated within specific social relationships ("What we owe to each other"). We describe how to implement such a model in artificial systems and discuss how such

* Corresponding Author. Email: clara.pretus@uab.cat

an implementation would allow these systems to (i) meet the needs of particular individuals based on human autonomy and fairness, (ii) be perceived as moral agents, reducing perceived dehumanization, and possibly enabling people to develop higher trust in AI.

2 Contractualism can help address human autonomy and fairness concerns in artificial systems

Contractualism, similarly to utilitarianism, is an impartial moral theory. That is, it seeks to provide a universal standard of morality that applies to everyone equally, and thus, it does not favor some individuals (such as family members) over others (such as strangers). This does mean that individuals should treat everyone else exactly the same. Instead, agents should rationally agree on which norms should govern the relationship to one another on the basis of mutual respect. Thus, the universal aspect of contractualism lies in the principle of mutual respect and rational agreement between agents rather than the specific content of norms that rule a given relationship. As opposed to utilitarianism, which focuses on the outcomes of an action, contractualism evaluates the action per se. For example, a contractualist agent understands it is wrong to punch somebody even if they miss. In this respect, it is similar to deontology, because it evaluates a given action against a set of norms. In contrast to deontology, this set of norms is dynamic and context-sensitive. In particular, contractualists deem an action appropriate when it can be justified based on the terms of a given relationship. The terms of each relationship, however, can evolve based on agreement among equals, but can never transgress the basic norm of mutual respect. Thereby, contractualism offers a flexible ethical framework grounded on social relationships.

One of the key strengths of contractualism is that people can deem an action inappropriate not only on the basis of pain and pleasure (like utilitarianism) but because of personal reasons. Because reasons can be responsive to the situation of others, people's preferences can incorporate not only their own complaints about a situation but also the complaints of others. By making moral agents responsive to other people's needs, contractualism attributes psychological capacities to moral agents such as reasoning and theory of mind – the ability to interpret other people's mental states. These capacities go beyond the capacity of feeling pain and pleasure. Thus, entities that are not perceived as capable of reasoning, such as non-human animals, are less likely to be considered moral agents. Meanwhile, contractualism offers the possibility for machines to be regarded as moral agents, as they may be able to reason about others' mental states sooner than they are to experience pain.

Precisely because moral agents are in tune with each others' needs, contractualism is particularly suitable to address current concerns over human autonomy and fairness in artificial systems. First, instead of limiting human desires to their own benefit or well-being, contractualism is sensitive to other-centered reasons. For example, a parent may prefer to reduce their own well-being by not getting enough sleep to take care of their child. These preferences may be difficult to model from a purely hedonistic perspective. Because contractualism allows for complex personal preferences, it is well-equipped for supporting human autonomy. Second, while utilitarians conceptualize fairness solely based on equal treatment, contractualists also attempt to give others what they are entitled to (starting with respect) based on the terms of their relationship. This introduces a minimal notion of accountability, as moral agents are bound to certain obligations towards others. Thereby, contractualism provides a framework to support both human autonomy and fairness in meaningful ways.

In contractualism, individuals contain multitudes. A majority may give up on a small benefit if it imposes a burden on just one person. Therefore, contractualists argue that they don't need aggregation. In a way, they are insensitive to quantity. This poses challenges to "saving one versus saving five" type of moral dilemmas, which contractualist solve in intricate ways (see Consensus-like aggregation section). In relation to this, while utilitarianism can be applied to all aspects of morality, contractualism is restricted to the domain of obligations among individuals. These constraints make it hard for contractualists to deal with non-human animals and future people, with which they cannot arrange a set of rights and obligations that will be reciprocated. The question arises as to whether these arrangements can be established in human-robot interactions. If artificial systems develop the status of moral agents based on their reasoning and theory of mind skills, then contractualism could be particularly useful in modeling human-robot interactions.

Overall, contractualism is an impartial ethical framework that evaluates the appropriateness of actions based on whether they can be justified to others. Justification can be based on self and other-centered reasons, allegedly making aggregation redundant. These properties make it suitable for supporting human autonomy and fairness in artificial systems. Of note, contractualism can only be applied in the domain of obligations among moral agents, and so artificial systems should complement it with additional ethical frameworks such as utilitarianism, which are more straightforward in dealing with quantities. Contractualism could be particularly useful for artificial systems that interact with humans and could help regulate human-robot interactions.

3 Can artificial systems be considered moral agents?

3.1 The moral status of artificial agents

To build contractualist artificial systems, they should first qualify as moral agents and moral patients. Moral agency refers to the capacity and responsibility of individuals to make moral judgments and choices and to act upon them. It encompasses distinguishing right from wrong, understanding the consequences of one's actions, and exercising free will in making ethical decisions [20]. Moral agency recognizes that individuals possess a moral compass and are capable of deliberating actions that can significantly impact themselves and others [19]. Moral agency also entails being accountable for one's actions and accepting the moral consequences that arise from them. It is a fundamental aspect of human nature that enables us to navigate the complex moral landscape and contribute to improving ourselves and society. An artificial moral agent should thus be able to predict and respond to the consequences of their actions. In contractualist terms, artificial moral agents should be subject to obligations toward others with which they establish a relationship.

A second aspect of the moral status of a system is moral patiency. Moral patiency characterizes individuals as objects of interests, rights, autonomy, or inherent value that should be recognized and respected by moral agents [3, 19, 11, 12]. For artificial systems to be moral patients, they should be regarded as holders of inherent value and rights. From a contractualist perspective, it would mean that humans that establish relationships with artificial systems would

be subject to obligations toward them. Because it is unclear whether artificial systems have an inherent value, we will focus on artificial moral agency in this manuscript.

3.2 Artificial systems as moral agents

If the definition of moral agency implies responsibility and a certain degree of awareness of the consequences of an action, can, and should, we attribute such moral status to artificial agents? Several authors proposed different accounts and answers to such questions [12]. For example, [4] argue that we should create a computational model of human ethics and implement this architecture so that the artificial system can reason autonomously, leaving aside the emotional involvement. Other authors, such as [21], argue that the lack of consciousness, mental states, and intentions does not allow for ascribing moral agency to artificial systems. Taking a less strict approach, one could argue that, even though artificial systems cannot fulfill the requirements to be considered full moral agents, they can still fall in the spectrum of moral agency.

Although the debate on whether artificial systems have or will have internal states such as beliefs, and desires, or will ever be conscious systems at all is still unresolved, we know from the literature that people do tend to ascribe such internal states to them. In other words, literature reports that humans tend to adopt an intentional stance toward embodied and disembodied artificial systems [13, 24, 23, 1, 30, 26]. Since literature on human-human interaction has reported correlations between the attributed level of intentionality and the attribution of moral responsibility to the consequences of humans' actions [7, 25, 27, 18], it is plausible to assume that adoption of intentional stance might entail attribution of moral agency. Moreover, literature in the field of human-robot or human-agent interaction showed that humans make attribution and have expectations about the moral status of artificial systems [5, 6, 22].

Thus, it is pivotal to integrate the classical approach to ethics of technology, focusing on humans' perception of such agents, rather than only questioning the moral status of such systems per se. Along the same lines, [16] and [29] argue that artificial agents do not need to have mental states or personhood to qualify as moral agents. It is sufficient that the exerted behavior has a degree of autonomy, interactivity, and adaptability to elicit the ascription of the status of a moral agent. Specifically, [16] introduced the idea of a "mindless morality". According to their account, a system needs to meet a series of requirements to be considered an agent: 1- the system needs to be able to change its states in response to a stimulus (interactivity); 2- the system needs to change a state without the prompt of a stimulus (autonomy); and 3- the system needs to change the rules according to which the states are modified (adaptability). In sum, Floridi and Sanders separate moral agency from moral responsibility, so that non-human systems can be considered (moral) agents but not morally responsible for their actions [16]. It is important to highlight that Floridi and Sanders' approach is non-anthropocentric, meaning that it does not analyze the systems according to the humans' perception of them. Their argument in favor of a non-anthropocentric view is that it allows for a non-dogmatic approach to answering the problem arising from the moral status of artificial systems.

A significant change in the direction of an anthropocentric approach to the problem was made by Mark Coeckelbergh in 2009 [10]. Coeckelbergh focuses on two main points: first, he argues that a non-anthropocentric approach ignores that for each non-human system, there is a human subject that defines the system as such. Indeed, technology itself is intrinsically bounded to humans, as we forged it. Thus, the lens applied to talk specifically about artificial systems needs to be anthropocentric. Secondly, following the first point, Coeckelbergh argues that, based on the perceived humans' experience of the interaction, we should ask "if, how and when this interaction and ascription could be justified" not questioning "what really goes on in there" [10]. That is in line with the literature on the investigation of the adoption of the intentional stance towards robots, for example, where the observer remains agnostic about whether the robot really has mental states. Nevertheless, the observer still interprets and explains the behavior of the robot as if the robot has them [23]. In summary, the approach proposed by Mark Coeckerbergh [10] relies on expanding the concept of which actors fall in the humans' social sphere, where the "other" with whom we interact can be a non-human agent. By adopting this frame, humans won't have to wonder about the "real" epistemological truth about the moral status of such agents, instead, they can adapt socio-cognitive mechanisms already in use in other (social) contexts [8, 28, 33, 31].

3.3 Guidelines for building a value-aware artificial intelligence

As previously discussed, artificial systems are likely to be considered as embedded in our social environments, and thus, we should ensure that these systems are built "for us", by adopting an anthropocentric perspective. This, according to us, is a necessary step in the development of artificial architecture (embodied and disembodied) because it will help to ensure that humans' autonomy in (moral) decision-making is not harmed. Artificial systems should, indeed, be a tool to empower humans, not only in their moral reasoning but also in their ability to exert their will.

In this context, it is important not to assume that the underlying principles of artificial systems are, by default, all the same. Instead, we should explicitly state and present the foundation of the models, so that transparency is always secured. Formosa and Ryan [17] explored the societal implications of perceiving artificial systems as social actors in the context of autonomous (moral) decision-making. She singled out three levels in which the systems can be of assistance to humans:

1. *more valuable ends*: by helping in defining goals or the needs to achieve them;
2. i*mproving autonomy competencies*: by indirectly assisting (i.e., taking over less relevant tasks). This will give humans more time for other relevant tasks. The systems can also assist directly the humans (i.e., a positive social interaction) during the decision-making process. Formosa highlights that in the case of direct assistance, the assistance needs to be positive and thus take into account all the factors that will lead to a satisfying interaction, such as social inclusion (Nash et al., 2018), leading to a reduced perceived anxiety level (Jeong et al., 2015; Pu et al., 2019);
3. *more authentic choices*: Walker and Mackenzie (Walker and Mackenzie, 2020) define a choice as authentic if one acts based on motives, desires, preferences, internal mental states, and reasons, to which we remain agnostic. Thus, via positive social interaction, these systems can help us by pointing out wrong values, norms, or biases, increasing humans' awareness of their moral reasoning.

Formosa also analyses the negative effects on human autonomy, in order to acknowledge the possible risks:

1. *fewer valuable ends*: when the systems autonomously select which information is relevant for the human (i.e., relevance not based on the human values);
2. *worse autonomy competencies*: "moral deskilling" of humans in moral decision-making [32], endowing the system to make the decision on our behalf;
3. *less authentic choices*: For example, by ascribing mental states to the behavior of the system, humans might feel as if the system would "watch" or "control" them, harming the authenticity of their choices.

Therefore, a (self-conscious) anthropocentric view in designing value-aware artificial systems, should result in a positive impact on our society [10, 15].

4 The Contract-Based Model of Moral Cognition for AI

4.1 General description of the model

We propose a model for AI inspired by contractualism, which could serve as a starting point for developing artificial systems that respect human autonomy and are perceived as moral agents. The Contract-Based Model is based on two main processes: (1) moral evaluation of behaviors based on how they affect specific recipients, and (2) moral learning based on human feedback. These two processes have been designed to prioritize human autonomy, as they allow the system to learn the personal preferences of the individuals who are affected by a given action and then evaluate behaviors based on this individualized criterion. In that sense, the model can be situated in the context of other value alignment solutions such as Cooperative Inverse Reinforcement Learning (Hadfield-Menell et al., 2016). These authors propose that robots must learn the reward function of a human by observing the human's behavior and interacting with them. By learning the human's reward function, the robot can align its values with those of the human and contribute to the maximization of value for the humans.

In our model, the preferences of the moral agent, that is, the entity that performs the action, do not affect how the action is evaluated. Only the preferences of moral patients are taken into account unless moral agents are also affected by the action and thus become moral patients as well. For instance, when someone steals food from a shop to avoid starvation, the shop owner is not the only moral patient. The person who steals the food is both a moral agent and a moral patient since not stealing imposes a large burden on them. Thus, the stakeholders involved in defining which norms are permissible are the moral patients of an action who hold contractualist relationships with the moral agent. Even the moral agent who steals has responsibilities to themselves as moral patients, such as keeping themselves alive. This arrangement circumvents the need for external stakeholders as long as rational agreement and mutual respect have been guaranteed. In the presented model, artificial systems only operate as moral agents, and so their preferences are not considered in moral evaluations.

The moral evaluation process aims to assess whether a behavior is appropriate or not for a given recipient by using moral norms. Thus, action recipient preferences are defined as weighted norms. Unlike other normative approaches, contractualism uses norms in a flexible way that allows different sets of norms to be relevant to different people in different social relationships. For instance, social hierarchy norms may be preferred by managers when interacting with workers but not when interacting with friends. As a result, the same action can be appropriate when applied to one individual in a given relationship but inappropriate for another individual and relationship. Thus, the model needs to be able to learn which norms are relevant for each individual in each type of relationship. The moral learning process is based on human feedback provided by action recipients. Therefore, people liking or disliking actions in a given relationship is the moral compass the Contract-Based model uses to learn how to evaluate future behaviors.

Box 1. Leveraging Large Language Models for Norm-Behavior Alignment and the Generalization of Norms
Researchers have started to explore the potential of large language modeling (LLMs) to solve norm and value alignment problems. For instance, Bakker et al. (2022) investigate how LLMs could help align diverse human preferences and find common ground. In this study, human participants provided opinions on moral and political issues. Human statements were then used as input for an LLM fine-tuned to generate consensus statements aimed at optimizing human approval ratings. The resulting consensus statements obtained a greater preference rate (70These capabilities could make LLMs suitable to support two critical functions in the Contract-Based model that require alignment: Norm-behavior alignment and the Generalization of norms.
1. Norm-Behavior Alignment: LLMs equipped with fine-tuning capabilities offer a valuable tool for computing the alignment between norms and behaviors based on linguistic features. Observed behaviors and collected human feedback can be transcribed and compared to internalized norms that are stored as text strings in the artificial system. The LLM could then process this input to assess the similarity of the behavior and norm statements and the alignment between them.
2. Generalization of Norms: LLMs can also facilitate the identification and merging of similar norms based on linguistic features. By inputting various norm descriptions, the LLM can analyze the linguistic characteristics and underlying principles, helping researchers identify norms that share commonalities. Similar to Bakker et al. (2021) this process could enable the generation of overarching statements by merging related norms.
Thus, LLMs could be used to support critical processes in the Contract-Based model of moral cognition for AI.

4.2 The moral evaluation process

The Contract-Based Model is a versatile tool that can be adapted to multiple use cases, and its function in the system is to inhibit "immoral" behaviors, filtering out behaviors that serve the system's purpose (e.g., giving directions in a hotel reception) but are inappropriate for a given recipient with which they hold a given relationship. In its current form, the model is suited to evaluate behaviors that affect one or more humans, which are defined as moral patients. Thus, after an action is selected by the system, the Contract-Based model activates only if the recipient is human. Once activated, two different processes support moral evaluation: 1) the computation of the alignment between the behavior and the norms stored in memory, and 2) the estimation of the relevance of these norms for the affected individual in the context of a relationship. The appropriateness of a given behavior results from combining these two streams. That is, a given behavior is deemed inappropriate if

it violates norms that are important for the recipient in the context of their relationship with the agent.

4.2.1 Norm-behavior alignment

To evaluate the appropriateness of a given behavior, only the norms that relate to the behavior will be relevant. Thus, an important step in the moral evaluation process is to assess which norms are related to the behavior by computing the alignment between different norms and the observed behavior. Norm-behavior alignment refers to the similarity between the norm and the actions involved in the behavior. The similarity between actions and norms can be computed based on the linguistic features of the description of the action and the description of the norm (see Box 1). For instance, the alignment with the norm "pull someone's hair" will be greater for the behavior "pull Sally's hair" than for "pull Sally's ear", although there will be some alignment with both of them. The Contract-Based model uses continuous values of alignment because complex behaviors may be related to multiple norms. For two norms to be different, their importance for different people should vary. If the importance of two norms across individuals is highly correlated, they are merged into a single norm (see Generalization of norms).

Box 2. Quantifying the relevance of norms for different individuals

The estimation of norm relevance is a quantitative problem that the Contract-based Model needs to address to conduct moral evaluations. In this process, the model estimates a value that corresponds to the relevance of each norm for a given individual. A possible way to do that is to use information about the individual's social group, the type of relationship they hold with the moral agent, and their specific id. The types of relationships between humans and robots are not as dynamic and variable as in human-human interactions. Thus, for the sake of simplicity, we will keep the type of relationship constant. The estimation of norm relevance is performed by a component of the system that uses the individual's id and their social group as input and produces a vector of weights in the range [-1, 1] as output, where -1 means maximal rejection of the norm and 1 maximal desirability. The length of the vector is equal to the number of norms stored in the system.

For each of the norms, the system's component samples from a mixture model, in which there is a normal probability distribution for each of the social groups. These probability distributions are univariate and bounded in the range [-1, 1], and each of them comprises all the members of the social group. The norm relevance value of the social group is the expected value computed in the standard maximization step of the Gaussian model. It is computed under the Gaussian distribution instead of the mixture one because the probability of belonging to the group (cluster) is always 1 and known a priori. This value can be used to evaluate new individuals for which the social group is known but their individual value is still unknown. These distributions are learned and change after the human feedback phase. Human feedback changes the mean of the distribution by assigning new values to the individual.

4.2.2 Estimation of norm relevance

Each person attributes different importance to different norms in different relationships. Thus, the Contract-Based model needs to first identify who are the recipients of the observed behavior. Recipients are characterized by at least two traits: their social group and the type of relationship they hold with the agent. The social group is used for computational efficiency while the type of relationship modulates the relevance of the norms. Computational efficiency is required when the artificial system selects actions that affect unknown individuals. In those cases, the social group offers a good trade-off between averaging for computational simplicity and allowing for cultural differences in moral preferences. The type of relationship between the agent and the recipient modulates how much different norms need to be taken into consideration. For instance, the obligations between an artificial agent and its owner can be different than those between the agent and a customer. These norms, in turn, may be different in different cultures and social groups. In sum, this step computes "weights" that express how relevant each norm is for the recipient of the behavior within their relationship with the agent.

4.2.3 The appropriateness score

Through the estimation of the norms' relevance and the computation of their alignment with the behavior, the Contract-Based model can assess the appropriateness of an action for a given recipient in a given relationship. If the appropriateness score is negative, the Contract-Based model will inhibit the behavior, and the action will not be selected. If it is nonnegative, there is no inhibition signal and the action can be selected. Nevertheless, there may be cases where the action is performed and the recipient is displeased. If so, the artificial system should first justify why the behavior was selected.

4.2.4 Justification

Because contractualist agents are driven by the desire to justify themselves to others, justification should be a core function of any contractualist artificial agent. In the Contract-Based model, justification is straightforward thanks to the human-centered design of the moral evaluation process. If the recipient provides negative feedback about a given behavior it means that one or more parameters in the moral evaluation processes need to be updated. Specifically, either none of the available norms was related to the behavior (norm-behavior alignment error) or one or more norms were related to the behavior but they were mistakenly considered irrelevant for the recipient in the evaluated relationship (estimation of norm relevance error). The latter could be related to the recipient holding norms that are different from their social group. The system can, thus, explain which of the two evaluation processes has failed and how they will update its parameters for future interactions with that recipient. The system's ability to explain its operations and update its parameters in response to human preferences places human autonomy at the center of the Contract-Based model.

4.2.5 The moral learning process

The Contract-Based model can learn and update its parameters for future moral evaluations using human feedback. When the recipient provides negative feedback, three sequential processes are triggered to enable moral learning: 1) generation of new norms, 2) generalization of norms, and 3) updating individual preferences. Processes 1) and 3) change future norm-behavior alignment, while 2) is related to norm relevance.

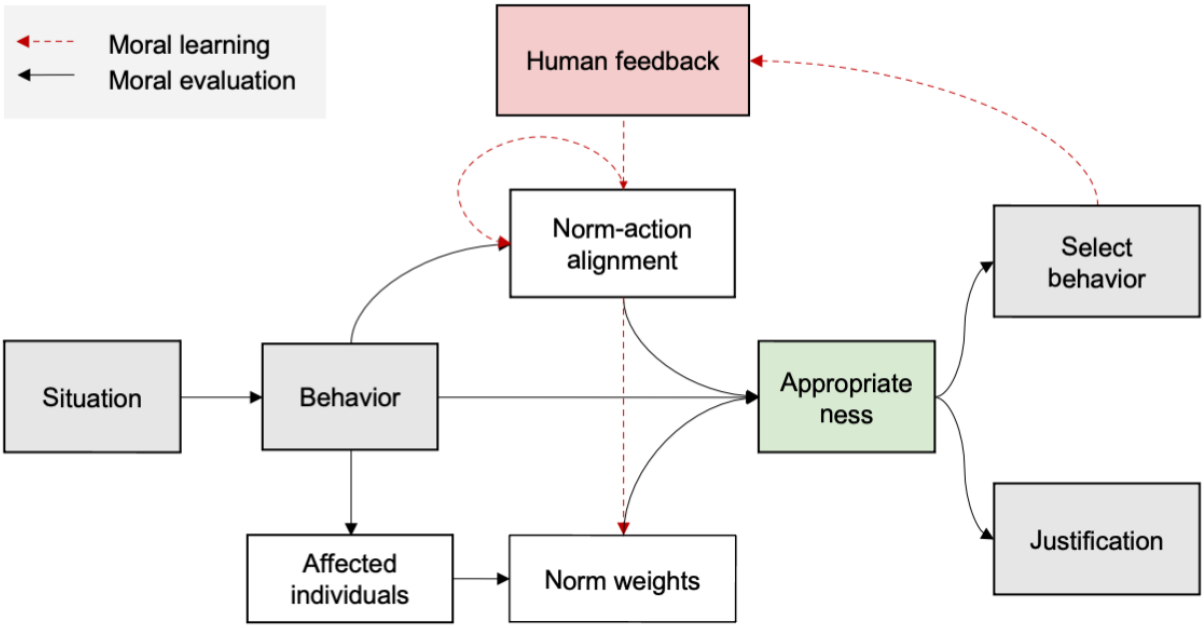

Figure 1. Contract-based Model of Moral Cognition for AI. The model computes an appropriateness score for actions embedded in a social situation based on the norms of the individuals affected by the action (moral evaluation). The model can refine the moral evaluation process through human feedback (moral learning).

4.2.6 Generation of new norms.

In case of a norm-behavior alignment error, the system did not properly identify the norms relevant to the observed behavior. This could be because none of the norms available in the system's memory properly aligned with the behavior. This caused the system to apply norms that are less related to the behavior. To fix this, new norms are needed. The first step in the generation of new norms is to compute the alignment between the behavior and all norms available in memory. If the low alignment between the behavior with all current norms is confirmed, the system removes the contextual features of the observed behavior's description, such as adverbs of time and place, and stores the simplified behavior description as a new norm. For instance, "pull Sally's hair at home" may be stored as "pull Sally's hair". Right after generating a new norm, the alignment between the new norm and the observed behavior is close to 1. However, through the generalization of norms (see next point), new norms are smoothed and turned into more abstract norms.

4.2.7 Generalization of norms

The purpose of this process is to keep the model as simple as possible for computational efficiency. For that, the system merges norms that have similar relevance across all registered individuals into more general norms. For that, the system assesses the dimensionality of the norms' relevance for all individuals. Then, if two norms are found to have a common pattern of relevance, a linguistic analysis can be conducted to merge them into a single norm, which will contain the common linguistic features of both norms.

4.2.8 Updating individual preferences

This process is triggered in two cases: (i) after generating new norms that better align with the behavior, and (ii) if existing norms already aligned well with the behavior but there was an error in the estimation of norm relevance for the recipient (e.g. because the recipient has very different preferences compared to their social group). In both cases, the system needs to update the relevance of the norms aligned with the behavior for that particular recipient. The extent to which the system updates the relevance of a norm depends on how aligned it is with the behavior for which the feedback has been provided. Individual human feedback feeds into the system and modifies the average relevance of norms within the social group the individual belongs to. Specifically, individual feedback modifies the relevance of norms for the type of relationship they hold with the agent in that social group. Thereby, human feedback by a given action recipient directly modulates future interactions with that recipient and with new recipients in the same social group.

5 Use of the model

5.1 Human-robot interactions

The Contract-Based Model of Moral Cognition for AI is a flexible framework that can learn human preferences and inhibit inappropriate actions based on these preferences. In new interactions, the artificial agent estimates human preferences based on the human's social group and the type of relationship they hold with them (e.g., the human is a customer). These capabilities are suitable for artificial systems that participate in social exchanges with humans, such as chatbots, domestic robots, or robots with customer-oriented jobs. In these settings, robots are in a social relationship with humans and humans expect them to display socially appropriate behavior. The Contract-Based Model allows artificial systems to meet these social expectations.

5.2 Assisting human-human interactions

The presented model could also be used to build moral enhancers, artificial systems that assist humans in moral decision-making in human-to-human interactions. Such moral enhancers could help humans overcome some of their cognitive limitations, such as memory capacity (not remembering people's preferences), assuming similar preferences to one's own (even when dealing with people from different cultures), or not being sensitive enough to the social context (treating work employees like friends). The Contract-Based moral enhancer could help humans keep these norms online for different relationships and individuals. Thus, the Contract-Based model could enable artificial systems to both behave as moral agents in human-robot interactions and function as assistants in human-human interactions.

Moral enhancers present of course a host of ethical problems that should be considered. For instance, moral enhancers may compromise human autonomy by promoting a standardized moral code and influencing people's moral choices in ways they have not consciously consented to. Determining responsibility for actions influenced by moral enhancers is also complex, posing challenges to accountability. In the long term, they could have unintended consequences, such as exacerbating social inequalities due to unequal access or making people overreliant on technological solutions and disincentivizing personal reflection. Therefore, moral enhancer technology should be accompanied by thoughtful discussion and careful regulation.

5.3 Artificial systems as moral patients

So far, we have discussed how artificial systems could function as moral agents, and behave in ways that are aligned with human preferences. This is only one dimension of contractualism: the idea that moral agents are bound to obligations towards others. The other half of a contract among equal parties involves that individuals on both sides have rights as moral patients. Because it is unclear whether artificial systems have an inherent value (see Section 3), it is also unclear whether they should have rights that humans should enforce. Thus, the Contract-Based model we present only operationalizes the terms that benefit humans and disregards any norms or "preferences" artificial systems may have. This may change in the future, as human relationships with artificial systems evolve. For instance, some humans verbally abuse virtual assistants such as Siri and Alexa (Harvard Business Review, 2016). While these types of behaviors are not damaging to the artificial system, they could be damaging to humans themselves, for instance, reinforcing and amplifying hostile attitudes and prejudices. Thus, the rights and obligations between humans and machines may be related in complex ways and future work should take this into consideration.

6 Conclusion

We propose a Contract-Based model of moral cognition that regards artificial systems as relational systems that are subject to a social contract. To articulate this social contract, we draw from contractualism, an impartial ethical framework that evaluates the appropriateness of actions based on whether they can be justified to others. In its current form, the Contract-based model characterizes artificial systems as moral agents bound to obligations towards humans. Specifically, this model allows artificial systems to make moral evaluations by estimating the relevance each affected individual assigns to the norms transgressed by an action. It can also learn from human feedback, which is used to generate new norms and update the relevance of different norms in different social groups and types of relationships. The model's ability to justify their choices to humans, together with the central role of human feedback in moral evaluation and learning, makes this model suitable for supporting human autonomy and fairness in human-robot interactions. Moreover, it can support artificial assistants that could function as moral enhancers to help humans navigate moral decision-making in human-human interactions. The Contract-Based model provides artificial agents a minimal ability to reason over people's mental states by computing moral evaluations based on human preferences. As human relationships with artificial agents evolve, the Contract-Based model could also incorporate new terms in the social contract between humans and machines, including terms that confer artificial agents a status as moral patients.

References

[1] Ahmad M Abu-Akel, Ian A Apperly, Stephen J Wood, and Peter C Hansen, 'Re-imaging the intentional stance', *Proceedings of the Royal Society B*, **287**(1925), 20200244, (2020).

[2] Matthew C Altman, 'A consequentialist argument for considering age in triage decisions during the coronavirus pandemic', *Bioethics*, **35**(4), 356–365, (2021).

[3] David Leech Anderson, 'Machine intentionality, the moral status of machines, and the composition problem', in *Philosophy and theory of artificial intelligence*, 321–333, Springer, (2013).

[4] Michael Anderson and Susan Leigh Anderson, *Machine ethics*, Cambridge University Press, 2011.

[5] Jaime Banks, 'Good robots, bad robots: Morally valenced behavior effects on perceived mind, morality, and trust', *International Journal of Social Robotics*, **13**(8), 2021–2038, (2021).

[6] Jaime Banks and Nicholas David Bowman, 'Perceived moral patiency of social robots: Explication and scale development', *International Journal of Social Robotics*, **15**(1), 101–113, (2023).

[7] Yochanan E Bigman, Adam Waytz, Ron Alterovitz, and Kurt Gray, 'Holding robots responsible: The elements of machine morality', *Trends in cognitive sciences*, **23**(5), 365–368, (2019).

[8] Francesco Bossi, Cesco Willemse, Jacopo Cavazza, Serena Marchesi, Vittorio Murino, and Agnieszka Wykowska, 'The human brain reveals resting state activity patterns that are predictive of biases in attitudes toward robots', *Science robotics*, **5**(46), eabb6652, (2020).

[9] Dallas Card and Noah A Smith, 'On consequentialism and fairness', *Frontiers in Artificial Intelligence*, **3**, 34, (2020).

[10] Mark Coeckelbergh, 'Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents', *AI & society*, **24**(2), 181–189, (2009).

[11] Mark Coeckelbergh, 'Robot rights? towards a social-relational justification of moral consideration', *Ethics and information technology*, **12**, 209–221, (2010).

[12] Mark Coeckelbergh, *Robot ethics*, MIT Press, 2022.

[13] Daniel C Dennett, *The intentional stance*, MIT press, 1989.

[14] Berkeley J Dietvorst and Daniel M Bartels, 'Consumers object to algorithms making morally relevant tradeoffs because of algorithms' consequentialist decision strategies', *Journal of Consumer Psychology*, **32**(3), 406–424, (2022).

[15] Frank Dignum, 'Autonomous agents with norms', *Artificial intelligence and law*, **7**, 69–79, (1999).

[16] Luciano Floridi and Jeff W Sanders, 'On the morality of artificial agents', *Minds and machines*, **14**, 349–379, (2004).

[17] Paul Formosa and Malcolm Ryan, 'Making moral machines: why we need artificial moral agents', *AI & society*, **36**, 839–851, (2021).

[18] Kurt Gray, Adam Waytz, and Liane Young, 'The moral dyad: A fundamental template unifying moral judgment', *Psychological Inquiry*, **23**(2), 206–215, (2012).

[19] Kurt Gray and Daniel M Wegner, 'Moral typecasting: divergent perceptions of moral agents and moral patients.', *Journal of personality and social psychology*, **96**(3), 505, (2009).

[20] David J Gunkel, 'The other question: can and should robots have rights?', *Ethics and Information Technology*, **20**, 87–99, (2018).

[21] Deborah G Johnson, 'Technology with no human responsibility?', *Journal of Business Ethics*, **127**(4), 707–715, (2015).

[22] Nils Köbis, Jean-François Bonnefon, and Iyad Rahwan, 'Bad machines corrupt good morals', *Nature Human Behaviour*, **5**(6), 679–685, (2021).

[23] Serena Marchesi, Davide De Tommaso, Jairo Perez-Osorio, and Agnieszka Wykowska, 'Belief in sharing the same phenomenological experience increases the likelihood of adopting the intentional stance toward a humanoid robot', (2022).

[24] Serena Marchesi, Davide Ghiglino, Francesca Ciardo, Jairo Perez-Osorio, Ebru Baykara, and Agnieszka Wykowska, 'Do we adopt the intentional stance toward humanoid robots?', *Frontiers in psychology*, **10**, 450, (2019).

[25] Andrew E Monroe and Bertram F Malle, 'Two paths to blame: Intentionality directs moral information processing along two distinct tracks.', *Journal of Experimental Psychology: General*, **146**(1), 123, (2017).

[26] Veronika Petrovych, Sam Thellman, and Tom Ziemke, 'Human interpretation of goal-directed autonomous car behavior', in *The 40th annualcognitive science society meeting, Madison, Wisconsin, USA, July 25-28*, pp. 2235–2240. Cognitive Science Society, (2018).

[27] Azim F Shariff, Joshua D Greene, Johan C Karremans, Jamie B Luguri, Cory J Clark, Jonathan W Schooler, Roy F Baumeister, and Kathleen D Vohs, 'Free will and punishment: A mechanistic view of human nature reduces retribution', *Psychological science*, **25**(8), 1563–1570, (2014).

[28] R Nathan Spreng and Jessica R Andrews-Hanna, 'The default network and social cognition', *Brain mapping: An encyclopedic reference*, **1316**, 165–169, (2015).

[29] John Sullins, 'Information technology and moral values', (2012).

[30] Sam Thellman, Maartje de Graaf, and Tom Ziemke, 'Mental state attribution to robots: A systematic review of conceptions, methods, and findings', *ACM Transactions on Human-Robot Interaction (THRI)*, **11**(4), 1–51, (2022).

[31] Esmeralda G Urquiza-Haas and Kurt Kotrschal, 'The mind behind anthropomorphic thinking: Attribution of mental states to other species', *Animal behaviour*, **109**, 167–176, (2015).

[32] Shannon Vallor, 'Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character', *Philosophy & Technology*, **28**, 107–124, (2015).

[33] Adam Waytz, John Cacioppo, and Nicholas Epley, 'Who sees human? the stability and importance of individual differences in anthropomorphism', *Perspectives on Psychological Science*, **5**(3), 219–232, (2010).