ARTIFICIAL EMOTIONAL INTELLIGENCE: INTEGRATING IRRATIONALITY IN MORAL RATIONAL AGENTS

Jacopo P. Gargano
Department of Computer Science
Politecnico di Milano
Milano, Italy
jacopopio.gargano@mail.polimi.it

ABSTRACT

Artificial Emotional Intelligence measures, understands, simulates and reacts to human emotions. This work aims to inquire about the possibility of integrating irrationality, intended as self-contradiction due to emotional instability or cognitive deficiency, into moral rational agents without making them irrational. Allowing agents to rationally act irrationally is important to avoid sub-optimal, unpredictable and undesirable behavior, potentially allowing for harm and catastrophic consequences. We first define moral rational agents as those agents always choosing the action with the greatest utility, while satisfying the properties of interactivity, autonomy and adaptability. We explore the various definitions of irrationality and identify the aforementioned one. We show how to integrate irrationality by expanding agents' utility calculation and preferences introducing sub-utility calculations, internal preference coefficients, and a distinction between logical and emotional preference. Furthermore, we show that in the proposed framework, which allows for irrational action, agents are still rational. Finally, we present the possibility of agents behaving randomly and unpredictably when not rational and discuss its consequences and possible solutions.

***K*eywords** Artificial Emotional Intelligence · Moral Agents · Irrational Decision-making

1 Introduction

Understanding, measuring, enhancing, and replicating *intelligence*, that is, the computational part of the ability to achieve goals in the world [McCarthy, 2004], has always been of great scientific interest. Artificial Intelligence (AI), coined by John McCarthy in 1955, is the science and engineering of making intelligent machines especially, in the form of computer programs, with the ultimate objective of allowing for the resolution of problems and the achievement of goals in the world as well as humans.

This work aims to address, both from a philosophical and technical perspective, the possibility of integrating irrationality, in the sense of self-contradiction (see Def. 2), in moral rational agents, as defined in [Floridi and Sanders, 2004], without making them irrational.

When dealing with problem solving in any kind of situation, namely taking an action after reasoning, machines, as well as humans, find themselves in a specific strategic setting, referred to as *game*. A game is a process consisting in: a set of agents[1], a set of *states*, including an initial situation and all the possible outcomes, a set of rules that agents must follow, the preferences of all the agents over the states, and a transition mechanism allowing the state of the game to change when an agent has taken an action [Maschler et al., 2013].

The reasoning process agents carry out to decide which action to take is based on their preference over the outcomes of the situation they are in. An agent is defined to be *rational* when it always takes the action has the highest expected outcome [Russell and Norvig, 2009]. The preferences of an agent depend on the specific virtues and goals that characterize it. Since in ordinary usage the words *moral* and *morality* have no precise and consistent use [Wallace and Walker, 2020],

[1] *Agents*: entities qualifying as the source of action, able to observe the surrounding environment through sensors and capable of reasoning.

we define an agent as *moral* when its behavior can be judged as right or wrong from an ethical perspective. For instance, both a human and a natural phenomenon such as a hurricane are agents, however only the former is considered as a moral agent.

Nowadays, AI is mostly focused on solving domain-specific problems, for instance, fraud detection, predictive maintenance, and recreational games. The complexity of these systems is already quite considerable and it does not allow for straightforward progress in building machines able to perform multiple uncorrelated tasks with significant performance. However, the definition of *intelligence* is domain-agnostic, referring to the concept of *general intelligence*: the ability to achieve complex goals in complex environments, adapting with insufficient knowledge and resources [Goertzel and Pennachin, 2007].

In 1905, with the aim of answering the question "Can machines think?", Alan Turing proposed the "Imitation Game" [Turing, 1950], where intelligence would be achieved by a computer program able to simulate human behavior in a text-based conversational interchange. However, through the years the goal of AI has shifted from developing an agent that simulates the behavior of a human to "*creating a nonhuman digital intelligent system, complementing human intelligence by carrying out data analysis and management tasks far beyond the capability of the human mind, cooperating with humans in a way that brings out the best aspects of both the human and the digital flavors of general intelligence*" [Goertzel and Pennachin, 2007].

If we want to include any kind of situation to the scope of an agent, especially those depending on the internal essence of the agent itself, we shall consider also feelings and discretion, shifting from *narrow* to *general* AI, focusing on Artificial Emotional Intelligence, which measures, understands, simulates and reacts to human emotions [Somers, Meredith, 2019], first analyzed in [Picard, 2000].

Our inquiry is hereby presented. In Section 2 we introduce the concept of *moral rational agents*, expanding on the definition provided by [Floridi and Sanders, 2004]. Section 3 explores the concept of *irrationality*, identifying the meaning we will refer to. In Section 4 we propose a framework to integrate logic and emotions in a generalized way into agents' preference. Finally, Section 5 explains how a moral rational agent integrating irrationality is still rational, w.r.t. the proposed framework. In Section 6 we briefly discuss why it is important that agents are rational, meaning they always choose the best action, and well-designed, leaving little room for stochastic behavior.

2 Moral Rational Agents

In every situation we can identify the entities that make it up. Some entities are *agents*, as they qualify as source of action; some are *patients*, being receivers of actions; some are both; and some others are neither. We will refer to entities that are both agents and patients simply as *agents*. The environment is the setting in which agents exist. It consists of all the entities, of its current state, and of rules. Rules define both the actions available to agents at a specific state of the environment, and the environment transition mechanism from a state to another for every possible action. For each state, each agent has its own preference over the actions performed by either itself or other agents. Since an action causes a change of state in the environment and being states identified also by the history of previous states and actions, we may generalize stating that agents have preferences over the environment's state. Their measure of preference is known as *utility*.

The choices an agent makes define its behavior. Morality is concerned with the principles allowing for the distinction between right and wrong behavior. It has been widely and deeply studied by *ethics*, which can be distinguished in three main branches [Bauer, 2020]:

- *Metaethics*, answering the question "How do moral values originate?";
- *Normative ethics*, investigating on what makes an action right or wrong and on the importance of consequences and intentions of actions;
- *Applied ethics*, focusing on determining whether or not specific actions in specific situations are morally justifiable.

In a multi-agent setting, as agents interact with one another, one's actions not only have effect on the agent that is the source of the action, but also on the other ones. To extend the scope of an agent, with the ultimate objective of reaching *General Intelligence*, the concept of morality must be included in agents. To embed morality in an agent, namely allowing it to consciously perform im/moral actions, we consider *moral agents*, embracing the definition of [Floridi and Sanders, 2004]:

Definition 1. *A* moral agent *is an agent satisfying the following three criteria*:

- Interactivity – *the agent and its environment (can) act upon each other.*

- Autonomy – *the agent is able to change state without direct response to interaction, performing internal transitions to change its state.*
- Adaptability – *the agent's interactions (can) change the transition rules by which it changes state.*

The first property allows agents to engage in a situation in which the actions of an agent influence the others. By being autonomous, an agent becomes independent from the environment it is in and it can focus on itself, diving into, possibly deep, domain-agnostic reasoning. Adaptability is the property of greatest interest for the scope of this work and it is based on the idea that agents have an internal state.

Each agent must in fact be different, especially if we wish to embed morality into it. If this were not the case, then all agents would make the same choices, as they would be guided by the same logical and ethical rules. Universal morality, based on Kantian deontology, would emphasize following strict duties, represented by maxims [Kant, 1785]. The categorical imperative would require having all agents morally evaluating a specific action in the same way, or at least quite similarly. However, this cannot be the case, as in practice there is no such thing as universally followed rules without contradiction. Moreover, this approach would limit discretion and feelings, leaving no room to *autonomy* and *adaptability*. Finally, if a *rational* agent were to follow maxims, then it might not be rational anymore. In fact, if an agent were to limit the available actions to those that respect universal maxims, then it might exclude the one action with the highest utility, as its preferences may clash with the maxims. Therefore, considering moral generalism, in which what is right is determined by applying ethical rules to situations, and moral particularism [Dancy et al., 2004], where what is right depends instead on the specific situation, we will adopt the latter.

Thus, each agent is defined by its internal state. This state collects the agent's characteristics in terms of morality, feelings, and logic. Moreover, it comprises the function used to compute the (expected) utility for a given state of the environment. This function depends on the agent's preferences, whose conception and values may change over time, as a moral rational agent is autonomous and adaptable (see Sect. 4).

3 Irrationality

"*Irrationality comprises a variety of psychological phenomena intermediate between error and madness*" is the beginning of Gardner's inquiry of irrationality [Gardner, 1993]. He suggests that one way of defining "irrational" would be to take it as the contrary of "rational". If we consider "rational" as in the sense of always taking the action with the highest expected utility, then defining "irrational" would be straightforward: an agent is irrational when it takes sub-optimal actions. However, we shall not limit to irrationality in this sense. If we integrated this definition of irrationality in moral rational agents, besides making them irrational, we would not be really representing intelligence as it is in practice, as it makes no sense for an agent to choose an action it considers as sub-optimal. Moreover, we would not be leveraging the properties of *moral* agents presented in Section 2, as we could discard morality and internal states and simply have agents choose actions randomly, resulting in sub-optimality.

We shall dig further into what makes behavior irrational. Gardner states "*the seeds of irrationality lie in a discrepancy between action and self-explanation, the recognition of which is bound up with the possibility of interrogation*" [Gardner, 1993]. According to him, once an agent has performed an action we may question it about its decision, making it reason once more. If there is inconsistency between the agent's answer and the action that was actually taken, then the agent "*is on the verge, at least, of being irrational*". This definition is closely related to the internal state transitions agents go through and also to the properties defining moral agents, especially that of autonomy.

Other definitions of irrationality can also be related to the internal state of agents. According to [Argenteri, 2006] and [Wikipedia contributors, 2019], irrationality manifests when an action is chosen through emotional distress or cognitive deficiency, or when an agent is dominated by passions.

Therefore, we consider the following generalized definition of irrationality, integrating the afore-reported definitions, for our inquiry:

Definition 2. Irrationality (*An* irrational action) *is acting (an action performed) in a state of emotional instability or cognitive deficiency, resulting in self-contradiction.*

The following are two examples of irrational behavior as we intend it throughout this paper, w.r.t. the aforementioned definition. It is irrational to get angry at a smartphone that is not working and smashing it on the ground. It is also irrational to cry when watching a sad movie. In both the first and the second example, the agent is in a state of emotional distress: in the first it may be angry, and in the second it is probably sad. The agent chooses the two actions as a rational, in the sense of best, response to its internal state (see Sect. 4). However, when the agent is back to a stable state, it would reason and admit that it would not perform the same action, giving more importance to logical reasoning. Namely

the agent would be aware of the fact that throwing a phone on the ground will not fix it, and that a movie may have a sad twist, but that it is just fiction and that it does not make much sense to cry. Nonetheless, the agent did perform an irrational action, but it did so rationally.

4 Preference and Utility

Moral rational agents are characterized by an internal state, which distinguishes them one from another. This state changes over time, as well as the way it changes does. The internal state of an agent includes its preferences P over the possible states of the environment, which are observable through sensors.

To measure preference, classical game-theoretical approaches have introduced the concept of *utility*. Traditionally, in most simple scenarios, utility $u(s)$ is a function of the state s of the environment only. However, in general, agents may calculate utility in different ways, depending on how they are designed, reason and evolve. This is why we need to generalize utility calculation to allow for more complex scenarios.

4.1 Sub-utility Functions

We consider utility as the weighted sum of several *sub-utilities*, computed through sub-utility functions $\bar{u}_p(s,t)$, each relative to a specific preference $p \in P$ of the agent over a specific state s of the environment at time t. Examples of preferences and relative sub-utilities are money, happiness, anger, satisfaction, and specific objects that are of one's possession, each considered over a state of the environment, namely a situation. It is important to note that, even in its classical definition, utility does not have a specific unit of measure. Indeed, one shall consider a way to generalize the measure of sub-utilities through dimensionless quantities. This is to say that preferences shall be pair-wise universally comparable (e.g., comparing happiness and money should be possible). Note that the comparison is the same universally, whereas the value given to each preference depends on the single agent (e.g., an agent may value happiness more than money). Here, we assume that comparing preferences is possible for all agents.

4.2 Internal Preference Coefficients

We introduce *internal preference coefficients* $k_p(t)$ to represent how valuable each preference p is to an agent at time t. The values of these coefficients vary agent by agent and depend both on the agent's nature, namely its initial design, and on its nurture, i.e. its experiences and evolution, considering the properties of moral rational agents introduced in Section 2. It is crucial to note that these coefficients define the internal state of the agent and do not depend on the specific state s of the environment. These coefficients define the agent's emotional state – its mood – allowing it to prefer preferences.

We distinguish preferences in two mutually exclusive categories: *logical* and *emotional*.

4.3 Logical Preference

We define a preference $p \in P_l \subset P$ as *logical* if it is based on unique logical reasoning, meaning that its corresponding sub-utility is computed in the same way by all agents meeting certain intellectual and volitional conditions for a specific state of the environment, at a specific time. This is to say that only agents in a compromised intellectual state – cognitive deficiency – would reason differently when calculating the associated utility.

Without loss of generality, we present a simple example of logical preference. With reference to a simple game[2], an example of a logical preference would be that over the outcome of the game. This preference is logical since the related sub-utility is uniquely computable by all intellectually stable agents applying the universally known and unique rules of the game. Any moral rational agent characterized only by this logical preference would want to win the game.

Generalizing, all moral rational non-intellectually-compromised agents able to experience only a logical preference will make the same choice regardless of their internal state. One may argue that if the agent gained something else by losing the game – as it is in the case of corruption – it would want to lose the game. However, if this were the case, we would have to include also the logical preference over money. In fact, considering both preferences, the agent would probably tend towards losing, depending on its internal preference coefficients.

[2] *Simple game*: a game in which agents either win or lose, that is, winning corresponds to $u = 1$, losing to $u = -1$. Note that a cognitively deficient agent would calculate the utility in a different way.

4.4 Emotional Preference

We define a preference $p \in P_e \subset P$ as *emotional* when its corresponding sub-utility is not uniquely computable. Sub-utility function $u_p(s,t)$ of emotional preference p varies agent by agent and changes over time as the agent exists, interacts with the environment, and evolves transitioning from an internal state to another, possibly modifying the way it does so. An example of emotional preference is anger. Not all moral agents get angry in the same way for a specific event. The level of anger an agent feels depends on the agent itself, on its nature and on its nurture.

Generalizing, emotional preference is a way to represent emotions in a moral rational agent.

4.5 General Framework

According to the presented framework, each agent is characterized by:

- the universal set of preferences $P = P_l \cup P_e$;
- $|P|$ functions $\bar{u}_p(s,t)$ to calculate sub-utilities, varying over time;
- internal preference coefficients $k_p(t)$, varying over time.

What is different among agents is the way sub-utilities are calculated – depending on initial design, experiences and evolution – and the way internal preference coefficients vary, that is, internal state transitions – depending on interactivity, autonomy, and adaptability. Note that one may argue that not all agents may be able to experience the same preferences or that they may only understand some preference as they "grow up". Then, one could simply set the sub-utility calculation to return 0, so that it won't contribute to the overall utility calculation, and eventually change it as the agent evolves.

The proposed utility function for an environment state s at time t is:

$$u(s,t) = \sum_{p \in P} k_p(t)\ \bar{u}_p(s,t) \tag{1}$$

Finally, note that the environment state s can be either the current environment state or the state the environment transitions to when taking an action a. The definition of $u(s',a,t)$ is the same as Equation 1, where s is the state the environment transitions to starting from state s' and taking action a.

5 Integrating Irrationality in Moral Rational Agents

Let us recall the definition of irrationality: acting in a state of emotional instability or cognitive deficiency, resulting in self-contradiction (Def. 2). We may finally investigate why the presented framework (Sect. 4) allows us to integrate irrationality in moral rational agents without making them irrational.

Cognitive deficiency is represented by a change in the sub-utility function $\bar{u}_p(s,t)$ of a preference p. Let us consider t_i as a time of cognitive deficiency or instability and t_s as one of stability. Then $\bar{u}_p(s,t_i) \neq \bar{u}_p(s,t_s)$. Note that the internal state coefficients do not necessarily change when an agent is cognitively deficient. Cognitive deficiency may refer both to logical and emotional preferences. An example for the former case is trivial: considering losing better than winning when playing a game. The latter is more interesting. Consider self-harm, this practice is universally associated with a very low utility, causing physical pain in humans and malfunctioning in machines. However, it seems that people do it to express their distress, or relieve unbearable tension [NHS UK, 2020], as it brings them relief. If we consider r as relief and s as the consequent state after taking a self-harmful action, then $\bar{u}_r(s,t_i) \gg \bar{u}_r(s,t_s)$.

The emotional state of an agent at time t is represented by the internal state coefficients $k_p(t)$, with $p \in P_e$. A state of emotional instability happens when there are some emotions strongly prevailing on others, especially negative ones. Without loss of generality, let us call $e \in P_e$ the prevailing emotion, and consider t_s and t_i as before. Then, we will have $k_e(t_i) \gg k_e(t_s)$. Note that $\bar{u}_e(s,t_i)$ is not necessarily different from $\bar{u}_e(s,t_s)$.

In both cognitive deficiency and emotional instability the expected utility u of the available actions at a certain state of the environment will be different at times t_s and t_i, that is, $u(s',a,t_s) \neq u(s',a,t_i)$. Therefore, if the agents we consider are moral rational agents, then at times t_s and t_i they will choose the action with the greatest utility. However, the two chosen actions may not coincide, resulting in self-contradiction, and therefore, irrationality.

Note that the considered class of agents is still rational, as they always choose the action with the highest expected utility, but they may experience irrationality in the form of self-contradiction when emotionally unstable or cognitively

deficient. This definitively shows that we are able to integrate irrationality in moral rational agents without making them irrational.

6 Why Should Artificial Agents Be Rational and Well-designed?

We now focus our attention specifically to artificial moral rational agents – as we cannot design humans, yet – and briefly discuss why it is important for them to be rational and well-designed.

Rationality in its classical game-theoretical definition is choosing the action with the highest (expected) utility among the available ones. It is very important that this property holds for moral rational agents. If agents were not rational, then they would choose sub-optimal actions, possibly resulting in stochastic behavior.

Stochastic behavior, especially in uncontrollable[3] artificial agents, is very undesirable. The properties of autonomy and adaptability of moral agents, as defined in Section 2, are the ones that allow for independence and evolution, which may lead towards unpredictable behavior, even for rational agents.

Independence, in the sense of inner thinking without stimuli from the environment, allows for spontaneous and unnoticeable transitions of the agent's internal state, as no particular event is needed to trigger a change. However, one may argue that the agent would still be following the transition rules embedded into it by the designer. This is true, unless the internal state it transitions to is one allowing it to wiggle out of its design. Therefore, autonomous agents' behavior, internal state shifts and inner reasoning are generally controllable.

Instead, evolution, in the sense of adaptability, is what can potentially make agents unpredictable. In fact, if agents are allowed to change the transitions rules of their internal state as they wish, they must be designed in such a way that does not allow for unwanted changes. For instance, if an agent is designed to seldom get angry, but as it evolves it tends to be more prone to getting angry, then the consequences may be catastrophic, especially if it can cause harm and if it is uncontrollable.

Therefore, it is very important for agents to be well-designed, allowing little or no room to stochastic changes, embedding moral values, and to be rational, so that the action they choose will always be the one respecting the most the principles they are designed by.

7 Conclusions

In this work we investigated the possibility of integrating irrationality, in the sense of self-contradiction due to a cognitive deficiency or emotional instability, in a moral rational agent without making it irrational. We first recalled the definition of rational and moral agent, referring to classical definitions and to [Floridi and Sanders, 2004]. Then, we explored the concept of irrationality, deriving the aforementioned definition to use throughout our inquiry. Moreover, we presented a way to allow for more complex and emotional agents, introducing preferences, sub-utility functions, internal preference coefficients, and distinguishing between logical and emotional preferences. Finally, through the proposed framework we showed moral rational agents can rationally act irrationally, and discussed the importance of them still being rational, well-designed, leaving little room for stochastic behavior.

In 1951, the American mathematician John Nash introduced the concept of *mixed-strategy Nash Equilibrium* (NE) [Nash, 1951]. A NE is a joint combination of strategies stable with respect to unilateral deviations, namely each agent has no incentive to deviate from its strategy given the strategies of the others, which behave rationally in the worst case. If we assume that agents' preferences and available actions are universally observable, then being able to compute NEs would allow agents to derive optimal strategies for every situation. This is the case, for instance, of recreational games such as poker [Brown and Sandholm, 2018]. Thus, one may wonder if living would still make sense, as everything may be precomputed and therefore known.

Nonetheless, in practice many phenomena are both unpredictable, stochastic and uncontrollable, such as natural events. Moreover, the assumption of knowing agents' preferences is very strong and not applicable to reality. Anyway, we may introduce ϵ*-optimality* allowing agents to choose one of the approximately optimal actions, controlling the permitted level of sub-optimality. Controlled randomness, instead, may be implemented in internal state transitions and in adaptability. Therefore, allowing for a bit of randomness and sub-optimality in the choice of actions would let artificial agents be more human-like and the aforementioned phenomenon of knowledge of the future would be contained, resulting in a more interesting existence.

[3] *Uncontrollable agents*: agents over which one cannot take control once deployed.

References

[Argenteri, 2006] Argenteri, S. (2006). Razionale e irrazionale. In *Treccani*. Istituto dell'Enciclopedia Italiana. Online; accessed 4-May-2020.

[Bauer, 2020] Bauer, W. A. (2020). Virtuous vs. utilitarian artificial moral agents. *AI & SOCIETY*, 35(1):263–271.

[Brown and Sandholm, 2018] Brown, N. and Sandholm, T. (2018). Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424.

[Dancy et al., 2004] Dancy, J. et al. (2004). *Ethics without principles*. Oxford University Press on Demand.

[Floridi and Sanders, 2004] Floridi, L. and Sanders, J. W. (2004). On the morality of artificial agents. *Minds and machines*, 14(3):349–379.

[Gardner, 1993] Gardner, S. (1993). *Irrationality and the Philosophy of Psychoanalysis*, pages 2–3. Cambridge University Press.

[Goertzel and Pennachin, 2007] Goertzel, B. and Pennachin, C. (2007). *Artificial general intelligence*, volume 2, page 73. Springer.

[Kant, 1785] Kant, I. (1785). *The moral law: Groundwork of the metaphysic of morals*.

[Maschler et al., 2013] Maschler, M., Solan, E., and Zamir, S. (2013). Game theory. *Cambridge University Press*.

[McCarthy, 2004] McCarthy, J. (2004). What is artificial intelligence?

[Nash, 1951] Nash, J. (1951). Non-cooperative games. *Annals of mathematics*, pages 286–295.

[NHS UK, 2020] NHS UK (2020). Self-harm. *National Health Service UK*. `https://www.nhs.uk/conditions/self-harm/`. Online; accessed 5-May-2020.

[Picard, 2000] Picard, R. W. (2000). *Affective computing*. MIT press.

[Russell and Norvig, 2009] Russell, S. and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, USA, 3rd edition.

[Somers, Meredith, 2019] Somers, Meredith (2019). Emotion ai, explained. *MIT Sloan*. `https://mitsloan.mit.edu/ideas-made-to-matter/emotion-ai-explained`. Online; accessed 6-May-2020.

[Turing, 1950] Turing, A. (1950). Computing machinery and intelligence-am turing. *Mind*, 59(236):433.

[Wallace and Walker, 2020] Wallace, G. and Walker, A. D. M. (2020). *The definition of morality*. Routledge.

[Wikipedia contributors, 2019] Wikipedia contributors (2019). Irrationality — Wikipedia, the free encyclopedia. Online; accessed 4-May-2020.