Moral Coppélia: Affective moral reasoning with twofold autonomy and a touch of personality

M.A. Pontier[1], G.A.M. Widdershoven[2], J.F. Hoorn[1], J.L. van Gelder[3] and R.E. de Vries[4]

[1] VU University Amsterdam, Center for Advanced Media Research Amsterdam, matthijspon@gmail.com
[2] VU University Medical Center, Amsterdam, The Netherlands
[3]Institute for the Study of Crime and Law Enforcement (NCSR)
[4]VU University Amsterdam, Faculty of Psychology and Education

Abstract. We present a moral reasoner, Moral Coppélia that combines rational ethical theory with affective states and personality traits. We, moreover, treat human autonomy in the sense of self-determination as well as making a meaningful choice. Our system combines connectionist bottom-up with utilitarian top-down approaches. Moral Coppélia can reproduce the verdicts of medical ethicists and health judges in real-life cases and can handle the emotional differences between logically identical problems such as the Trolley and Footbridge dilemma. It also deals with properties of character and personality such as honesty and humility to explain why logic reasoning is not always descriptive of actual human moral behavior. Apart from simulating known cases, we performed a split-half experiment with the responses of 153 participants in a criminal justice experiment. While fine-tuning the parameters to the first half of the data, the encompassing version of Moral Coppélia was capable of forecasting criminal decisions, leading to a better fit with the second half of the data than either of the loose component parts did. In other words, we found empirical support for the integral contribution of ratio, affect, and personality to moral decision making, which, additionally, could be acceptably simulated by our extended version of the Moral Coppélia system.

1 INTRODUCTION

The need for ethical machines

Increasingly, computer systems run our lives autonomously. They do so in matters of increasing importance. Just as an example, microwave communication deployed in stock market trading makes deals at the millisecond and breaks them again way beyond the speed limits of human mental chronometry. Algorithmic trading systems decide on financial transactions on their own authority; no human interference whatsoever. The velocity by which a fabulous number of interactions emerges, is an ideal cover for fraudulent behaviors and criminal conduct. Some might say that those developments are a threat to human autonomy; that we should stop building such technologies. The other option could be to make autonomous systems such that they can act in an ethical way. Create a moral stop button so to speak. That way, we may employ the advantages of performing faster, better, cheaper, and more reliably without having to fear a loss of autonomy. As Rosalind Picard [17] puts it: "The greater the freedom of a machine, the more it will need moral standards." Another, slower, approach would be that humans make the decisions while the system serves as an advisor – also on moral matters.

Moral decision making is arguably one of the most challenging tasks for computational approaches to higher-order cognition [29]. To contribute to a field that is variously known as Machine Morality, Machine Ethics, or Friendly AI (ibid.), we developed a moral reasoner that we applied to ethical dilemmas in healthcare and crime. The WHO [31] anticipates a lack of resources and healthcare personnel worldwide while progressively, robot systems are marshaled for care support. For example, Robins et al. [21] used mobile robots to treat autistic children. Paro [28], a robotic baby-seal that encourages positive mental effects and that is used at eldercare facilities for therapy with Alzheimer patients. In comparison with living dogs, the AIBO robot dog helped just as good to reduce the loneliness of elderly people [4]. Such robotic care interventions, however, should not impede the promotion of human values or compromise the dignity of patients at such a vulnerable and sensitive time in their lives [27]. Additionally, we opted for crime-related scruples, because human behavior typically is far from being morally ideal [1]. Perhaps that a moral reasoner could come to serve as the ethically 'better half' of a potential perpetrator.

Related work

There are multiple perspectives to take on moral issues, and picking one readily determines the type of system that will be developed. Approaches go from narrow-focused casuistry to wide-scoped utility for the world, sampling judgments from the bottom up to imposing principles from the top down. Next, we will provide an example of each and propose a synthesis thereafter.

Casuistry particularly looks at previous cases in which agreement is established about the 'correct' response – correctness of course, according to the moral principles or world view at play. When the machine is confronted with a new case, analysis of the similarities with the previous cases helps to formulate what the correct response would

be. For example, Guarini [12] offers an implementation of casuistry in which a neural network learns patterns of judgment from training examples of ethical dilemmas with a known 'correct' response. After learning, this system offers responses to new ethical dilemmas that may be deemed plausible. However, reclassification of cases remains problematic due to a lack of reflection and explicit representation. The conclusion, therefore, is that casuistry alone is insufficient.

At the other end of the spectrum, the ethics about duties maximizes the total amount of 'utility' (here, a measure of happiness or wellbeing), not just the one of a specific case. The 'big picture' view of moral principles is that ethics is about general duties and, on the flip side, the rights of individuals [29]. For instance, if one should kill one person to save five, killing the one person seems – in this case – to maximize the total amount of utility. After all, compared to the decision of inaction, this decision leads to a situation with four survivors [2]. In taking a broad perspective, however, the decision to kill any one person also makes it more acceptable to kill human beings in other cases (cf. passivism). Hence, inaction may be preferred. With the intuition never to kill (although many may be saved), it is likely that overall utility in the world will be higher.

An approach that uses judgments from the bottom up is demonstrated by Rzepka and Araki [22]. Their system learns to make ethical decisions based on Web-based knowledge, 'independent from the programmer.' They argue it may be safer to imitate millions of people than a handful of ethicists and programmers. Whereas this seems useful for simulating and describing human ethical behavior as is, this 'crowdsourcing' approach to morality may be questionable if the goal is more normative and directed at exhibiting exemplary behavior. After all, the system bases its decision on the average behavior of humans in general, misbehavior included.

Anderson and Anderson [3] agree with this view and address the need for top-down processes. The two most dominant top-down mechanisms they distinguish are (1) utilitarianism and (2) ethics about duties. Utilitarians claim that morality ultimately is about maximizing the total amount of utility in the world. For instance, Anderson, Anderson, and Armen [2] argue that the ideal ethical theory incorporates multiple moral duties with some sort of a decision procedure to determine the ethically correct action in cases where the duties give rise to conflicting advice. Their system learns rules from examples using a machine-learning technique. After learning, the system can produce correct responses to new, as yet unlearned, cases.

However, according to Wallach, Franklin and Allen [30], the model of Anderson, Anderson, and Armen [2] is rudimentary and cannot accommodate the complexity of human decision making. In an attempt to synthesize top-down and bottom-up moral decision faculties, these authors argue that the capacity for moral judgment in humans is a hybrid of bottom-up evolution and learning as well as top-down theory-driven reasoning. To handle diverse inputs, moral robots eventually should become dynamic and flexible through bottom-up systems, while choices and actions are subjected to top-down principles, representing desired ideals. Wallach, Franklin, and Allen [30] explored the possibility to implement moral reasoning in LIDA. This model of human cognition combines the collection of sensory data with making sense of a current situation to predict what the results of actions will be.

Our contribution also combines bottom-up structures with top-down knowledge of moral duties. It balances those duties to compute a level of morality that serves as an estimation of the influence on the total amount of utility in the world. Different from all other approaches, our focus is not on ethical reasoning alone; we mix in rational choice with affective concerns, increasing the fidelity of the simulation of human moral decision making, which after all is neither free from emotional shading nor from aspects of personality.

2 MORAL COPPELIA

In this section, we will construct our system from theory and published data. That effort will amount into a moral reasoner that takes affective states into account, has a sophisticated understanding of autonomy, and makes decisions in line with certain personality types. After that, we will do a number of simulations of moral decision making in healthcare, health law, and crime, building up the complexity of its considerations factor by factor.

Silicon Coppélia

According to Tronto [23], care is only thought of as good care when it is personalized. For artificial systems, being personal may be a challenge. Our point of departure is the way people and machines build up a relationship with one another through interaction, in which ethical appraisals play a central role. Hoorn, Pontier and Siddiqui [15] modeled the AI system Silicon Coppélia after the way users evaluate artificial others. Silicon Coppélia is capable of generating and regulating affective states and processes and that can simulate emotions in response to other agencies (i.e. humans). Most importantly for our current purposes, Silicon Coppélia has an affective decision-making module that can trade rational for affective choice [14], calculating the expected satisfaction of potential actions. "Involvement" and "distance" are the affective components during the decision-making, whereas "use intentions" (the willingness to continue interacting) and the general expected utility to achieve a goal represent the more rational aspects. The system contains a library of

goals and each agent has a level of ambition for each goal within the domain [-1, 1]. Expected utility for the agent is:

ExpectedUtility(Action, Goal) =
Belief(facilitates(Action, Goal)) * Ambition(Goal)

Given the level of ambition for a goal and the believed facilitation of that goal by an action, the agent calculates the expected utility for itself of performing that action regarding that goal by multiplying the believed facilitation of the goal with the level of ambition for that goal. Emotions such as hope, joy, and anger are generated using appraisal variables (e.g., the likelihood and achievement of goal-states). Each emotion has a desired value, which it tries to approach through achieving goal-states, including those accomplished through rational means. In other words, affective and rational forces are combined in the decision-making process.

Moral Reasoning

Pontier and Hoorn developed Moral Coppélia [18]. This system is capable of moral reasoning and deciding on ethical dilemmas in the same way as medical ethical professionals do. Moral Coppélia roots in the (standard) prioritization of ethical principles, which it sees as moral goals to achieve; in order of increasing importance: justice, beneficence, non-maleficence, and autonomy [5].

The prioritization of ethical principles, nevertheless, is not carved in stone. Anderson and Anderson [3] posit that a caretaker should challenge a patient's decision only if the patient is not capable of fully autonomous decision making (e.g., the patient has irrational fears about an operation) and if there is either a violation of the duty of non-maleficence (e.g., the patient is hurt) or a *severe* violation of the duty of beneficence (e.g., the patient rejects an operation that will strongly improve his or her quality of life). In other words, patient autonomy may be the most important duty but provided that the patient is capable of being fully autonomous. If not, other moral concerns come into play.

With this in mind, Moral Coppélia tries to maximize the total amount of utility – in a moral sense – for every agency involved in a situation so to satisfy everyone's needs as much as possible, selecting those actions that best serve the moral goals of all. Moral Coppélia calculates the estimated level of Morality of an action by taking the sum of the ambition levels of the four moral goals multiplied with the beliefs that the particular actions facilitate the corresponding moral goals. When moral goals are believed to be better facilitated by a moral action, the estimated level of Morality becomes higher.

As can be seen Figure 1, this can be represented as a weighted association network, where moral goals are associated with the possible actions via the belief strengths that these actions facilitate the four moral goals.

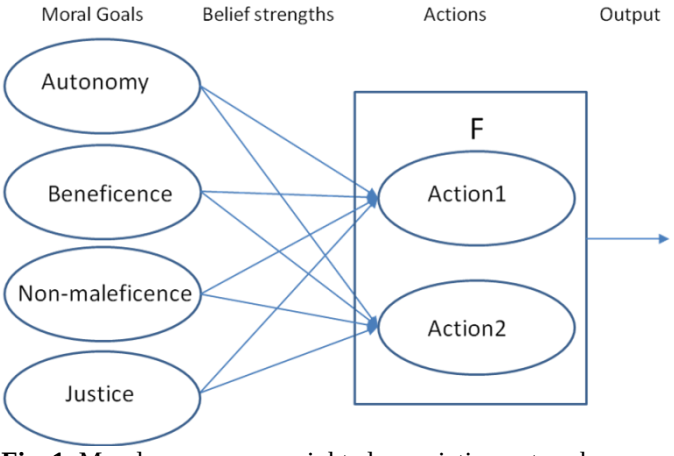

Fig. 1. Moral reasoner as weighted association network

The Morality of an action is estimated as follows:

Morality(Action) =
Σ_{Goal}(Belief(facilitates(Action, Goal)) * Ambition(Goal))

Note that this is similar to calculating the expected utility in Silicon Coppélia. The agent prefers actions with a high level of expected utility for itself. Additionally, it prefers actions with a high level of (rational) morality, which could be seen as expected utility for everyone. The more emotional influences consist of preferring actions with a positivity and negatively levels close to the levels of biased involvement, and biased distance, respectively. The biases account for individual defaults or "personality," being a positively or negatively oriented person.

Twofold autonomy

Oftentimes, autonomy is equated with self-determination. In this view, people are autonomous when they are not influenced by others. Beauchamp and Childress [5] are criticized for focusing too much on autonomous decisions [9]. Autonomy is not just being free from external constraints but can also be conceptualized as being able to make a meaningful choice, which fits in with one's life-plan [32]. In this view, a person is autonomous when s/he acts in line with well-considered preferences. This implies that the patient is able to reflect on fundamental values in life. Core aspects of autonomy as self-determination are mental and physical integrity and privacy. Central in autonomy as ability to make a meaningful choice is to have adequate information about the consequences of decision options, the cognitive capability to make deliberate decisions, and the ability to reflect on the values behind one's choices. Autonomy as self-determination can be called negative freedom, or 'being free *of*'. Autonomy as the ability to make a meaningful choice is called positive freedom or 'being free *to*' [6]. In this paper, we will use the notions of 'negative autonomy' and 'positive autonomy,' respectively, with a more complex

implementation of the moral principle of autonomy in Moral Coppélia [20].

The notion of positive autonomy is often used as an argument for not following people's immediate and often unhealthy wishes, and demanding more well-considered choices from the patient, which tend to be healthier. In this sense, positive autonomy may seem to come close to beneficence. Yet, autonomy as being able to make a well-considered choice is not the same as beneficence. Reflection on and deliberation about values can help people to behave in a more healthy way but this is not necessarily so. Reflection might result in people taking health risks in favor of other important values. An example is the conscious refusal by Jehovah's witnesses of blood transfusion.

In medical practice, sometimes the self-determination of the patient needs to be constrained on the short-term to achieve positive autonomy on the longer term. When a patient goes into rehab, his or her freedom can be restricted for a limited period of time to achieve better cognitive functioning and self-reflection in the future.

In our model, we divide autonomy into negative autonomy and positive autonomy. Negative autonomy can be seen as self-determination - or being free *of* others - and consists of the sub-principles physical integrity, mental integrity and privacy. Positive autonomy can be seen as the capability to make a deliberate decision – or being free *to* choose - and consists of having adequate information, being cognitively capable of making a deliberate decision and reflection. All variables in the model are represented by a value in the domain [0, 1].

To be autonomous, both the conditions for positive autonomy and negative autonomy are relevant. Ideally, both are present to a large extent. When self-determination is compromised, one is not able to make an autonomous decision, because this decision is made by others; the person is not free *of* others to make their own decision. When a person is not able to deliberate, the person is also not autonomous. The person may be free *of* others to make a decision, but not free *to* make an autonomous decision, because s/he lacks the ability to do so. This is reflected by:

Autonomy = Positive_autonomy * Negative_autonomy

When negative autonomy (or self-determination) is 0, autonomy will also be 0. When positive autonomy (or the capability to make a deliberate decision) is 0, autonomy will also be 0. For being autonomous, both negative autonomy and positive autonomy should be high.

Positive autonomy can be divided in having adequate information, cognitive functioning and reflection. For calculating positive autonomy from these three variables, we use the same reasoning as for calculating autonomy straight. Each should be present to some extent; the higher they are, the more autonomy is present. Without any information about the consequences of a decision, it does not matter whether one could have made a reasoned and deliberate decision while having this information. When one is mentally handicapped, it does not matter whether adequate information is available. When a decision is made without reflection, it does not matter whether one would have the cognitive capabilities and information to do so. The formula for calculating positive autonomy is similar to that for calculating autonomy.

Positive_Autonomy =
Information * Cognitive_Functioning * Reflection

When one of the three variables is 0, positive autonomy will also be 0. For being capable of making a well-reflected, deliberate decision, all conditions for positive autonomy need to be met to some extent.

Negative autonomy is divided into physical integrity, mental integrity, and privacy. For calculating negative autonomy, or self-determination, a different method is chosen. If privacy is constrained, but physical and mental integrity are left intact, the level of self-determination can be higher than the level of privacy alone. For calculating negative autonomy, a weighed sum of the three variables is taken, as can be seen here:

Negative autonomy =
w_p*Privacy + w_m *Mental_Integrity + w_{ph} *Physical_Integrity

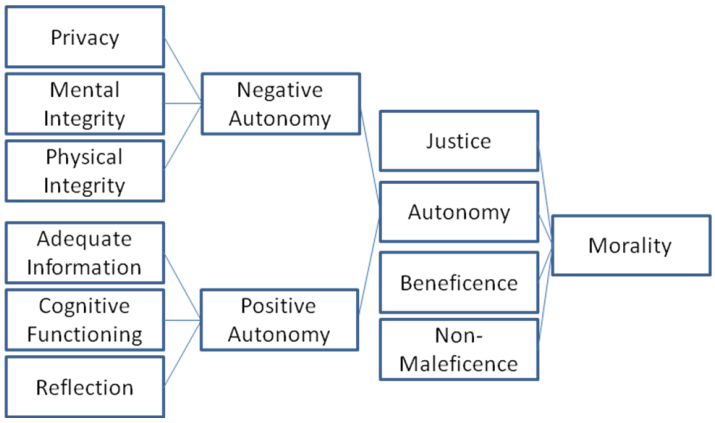

Fig. 2. Moral Coppélia with twofold approach autonomy

When making a decision that may influence the autonomy of a patient, the system will make an estimation of how each of the six variables will change. After doing so, it can calculate the resulting autonomy of the patient for every possible decision option and use the outcome to estimate how morally good or bad each decision option is. The calculated levels of both autonomies simply feed into 'autonomy' to establish the morality of each action (Figure 2).

Moral Coppélia calculates the estimated level of morality of an action by taking the sum of the ambition levels of the three moral principles of each type of autonomy multiplied by the beliefs that the particular actions facilitate the corresponding moral principles. When moral principles are believed to be better facilitated

by an action, the estimated level of morality will be higher, as in:

$$\text{Morality(Action)} = \sum_{\text{Goal}} (\text{Belief(facilitates(Action, Goal))} * \text{Ambition(Goal)})$$

With this rule, we conclude our discussion and implementation of the rational side of ethics. We will now turn to the affective side of Silicon Coppélia to see how it may interfere with the straightforward reasoning by Moral Coppélia.

Affective Moral Coppélia with twofold autonomy

Another criticism of the theory of Beauchamp and Childress [5] is that it is too fixated on rational argumentation and that social processes such as interpretation and communication are underexposed [16]. For decades, research on moral judgment has been dominated by rationalist models, in which moral judgment is thought to be motivated by moral reasoning. However, more recent research indicates moral reasoning is just one of the factors motivating moral judgment. According to Haidt [13], moral reasoning frequently is a *post hoc* construction, generated after a judgment has been reached.

Both reason and affect are likely to play important roles in moral judgment. Greene et al. [10] find that moral dilemmas vary systematically in the extent to which they engage emotional processing and that these variations in emotional engagement influence moral judgment. Their study was inspired by the difference between two variations on an ethical dilemma: the Trolley dilemma and the Footbridge dilemma. Both dilemmas are relevant to nursing practice, as we will explain next.

In the Trolley dilemma, a runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save them is to hit a switch that will turn the trolley onto an alternate set of tracks where it will kill one person instead of five. Are you to deflect the trolley to save five people at the expense of one? Most people would say "yes."

In the Footbridge dilemma, a trolley threatens to kill five people. You are standing next to a stranger on a footbridge that spans the tracks, in between the oncoming trolley and the five people. The only way to save the five people is to push the stranger off the bridge, onto the tracks below. He will die if you do this, but his body will keep the trolley from crashing into the others. Will you save the five others by pushing this stranger to his death? Most people say "no."

Care professionals face similar dilemmas, involving questions of fair distribution of resources among patients. If one patient needs to go and five others need to be fed, where will you allocate your time? Feeding five at the expense of one lying in his doings? The Trolley answer would be "yes." But what if that one patient is your little girl? The Footbridge frame would definitely say: "No!"

According to Greene et al. [10], there is no set of consistent, readily accessible moral principles that captures people's intuitions concerning what behavior is or is not appropriate in these and similar cases. In other words, the different human moral decision-making processes in the Trolley dilemma and the Footbridge dilemma (and similar dilemmas) cannot be explained by rational principles alone. Therefore, human moral decision making cannot be simulated in a moral reasoning system based on pure principlism.

Greene et al. [10] hypothesized that the crucial difference between the Trolley dilemma and the Footbridge dilemma lies in the latter's tendency to engage people's emotions in a way that the former does not. They suggested that the thought of pushing someone to his death is emotionally more salient than the thought of hitting a switch that will cause a trolley to produce similar consequences. We argue, however, that the difference is related to the fact that the person on the footbridge is standing close by, whereas the people on the railway track are positioned far away. And it is this emotional response that accounts for people's tendency to treat these cases differently.

The fMRI studies and behavioral results of Greene et al. [10] support this view. Moral-personal dilemmas (those relevantly similar to the Footbridge dilemma) engage emotional processing to a greater extent than moral-impersonal dilemmas (those relevantly similar to the Trolley dilemma), and these differences in emotional engagement affect people's judgments. In line with these findings, Wallach, Franklin and Allen [30] argue that even agents who adhere to a deontological ethic or who are utilitarians may require emotional intuition as well as other ''supra-rational'' faculties, such as a sense of self and a theory of mind.

Therefore, we decided to let the affective side of Silicon Coppélia impact the rational and principled line of argumentation by Moral Coppélia. Moreover, within the Moral Coppélia sub system, we split up the ethical principle of autonomy into a positive and a negative variant (see previous section). We defined a new variable called *ExpectedEmotionalStateAffect* (EESA) within the domain [0, 1]. A high EESA indicates that an action is expected to improve the emotional state of the agent (*Positive State Affect*), whereas a low EESA indicates that an action is expected to worsen the emotional state (*Negative State Affect*).

We also sophisticated certain aspects of the affect side of Silicon Coppélia. Previously, emotions were regulated implicitly, selecting actions that lead to desired goals. Because the earlier Silicon Coppélia assumed that emotions arose from beliefs about goal-states and actions were related to goals, actions were automatically selected that were thought to lead to desired goal-states and accompanying desired emotions. However, no beliefs

about how actions directly relate to a desired emotional state existed in the system. Therefore, we added the emotion regulation strategy called *situation selection* [11] as proposed by Bosse, Pontier, and Treur [7]. For calculating the EESA, we defined the variable *ActionEmotionBeliefs* (AEB). An AEB(action, emotion) represents the belief that an action will lead to a certain level of emotion. For example, an AEB(shoplifting, excitement) of 0.6 represents the belief that shoplifting will lead to a level of excitement of 0.6. The ExpectedEmotion that follows from this belief is calculated according to:

ExpectedEmotion =

$(1-\beta) * \text{AEB(action, emotion)} + \beta * \text{current\_emotion}$

Here, the persistency factor β is the proportion of emotion that is taken into account to determine the ExpectedEmotion. The new contribution to the emotional response level is determined by taking the appropriate AEB. To determine the EESA of an action, a weighed sum of the discrepancy between desired emotions and expected emotions after performing the action is subtracted from 1. For the sake of simplicity, the weights *w(i)* can be set to the same level for all emotions added to the system:

$\text{EESA(action)} = 1 - (\sum_{n}^{0} \boldsymbol{W}(i) * (\text{Desired(emotion(i))} - \text{ExpectedEmotion(action, i)}))$

To determine the expected satisfaction of a choice, a weighed sum is composed of the Morality variable, the rational ExpectedUtility, and the emotional EESA of the action:

ExpectedSatisfaction(Action) =
w_{mor} * Morality(action) +
w_{eu} * ExpectedUtility +
w_{emo} * ExpectedEmotionalStateAffect

This way, the agent system derives satisfaction from a configuration of actions with a high level of expected utility for itself; from actions with a high level of morality, serving the expected utility for all, and from actions that are believed to lead to a desired emotional state. In the next section, we will turn to the impact of personality on moral judgment, the literature of which is drawn from criminology.

HEXACO personality

Rational choice theories of criminal decision making assume a reasoning agency that balances costs against benefits. According to rational choice theories, people will offend when they expect that potential benefits exceed the anticipated costs; they will refrain from offence when the balance is reversed. Rational choice and deterrence models do not specify the psychological mechanisms according to which criminal decision making works [26]. Consequently, they do not specify how emotions such as anger, fear, or defiance of the offender influence the criminal calculus and alter risk concerns.

The emotion and rational view are not distinctively separate and may actually complement one another. Hence, both perspectives should be included in models that attempt to explain crime [25]. The interplay between cognition and affect was always prominent in dual-process theories of information. When people make judgments and decisions or engage in problem solving, two partially independent but mutually influential information processes are operative [26].

Van Gelder [24] argues that criminal decisions also invoke these two process types. In this hot-and-cool approach, the cool cognitive mode is sensitive to risky outcomes and risk probabilities. Therefore, it responds to severity of the sanction and ascertains certainty, as suggested by deterrence theorists. This mode also balances costs against benefits and ponders the long-term repercussions of criminal conduct. It functions according to the logics assumed by rational choice theory. The affective mode remains largely unresponsive to such probabilities (e.g., [25]) and illuminates why severe punishment has modest or no effect on crime rates and why recidivism rates are as high as ever. Particularly in crimes out of passion, short-term considerations outweigh the long-term consequences. Focus is on the here-and-now; the choice is for immediate benefits. This hot-and-cool approach is incorporated by the so called HEXACO model of criminal personality. In predicting criminal behavior, the relationship between personality, ratio, and affect outperforms existing measures of self-control and personality, such as the well-known Big Five model.

HEXACO works from the dimensions of Extraversion, Conscientiousness, and Openness to Experience, while Agreeableness and Emotionality are rotational variants of Agreeableness and Neuroticism. What makes HEXACO really different is the *Honesty–Humility* dimension, which seems to hold across cultures. It refers to being interpersonally genuine, not taking advantage of others, aversive of fraud and corruption, uninterested in status and wealth, being modest and unassuming. Van Gelder and De Vries [25] suggest that HEXACO (i.e. Honesty-Humility) is a strong predictor of criminal behavior throughout.

To the best of our knowledge, no computational models exist that include rational as well as affect and personality aspects in predicting crime. Therefore, we related Moral Coppélia's rational achievement of goals and the resulting ExpectedUtility of an action to a measure of Perceived Risk. Additionally, beliefs about emotional consequences of actions (EESA) were related to Negative State Affect, while the weight of Morality was related to the value of 'Honesty-Humility.' To calculate the Expected Satisfaction of a choice, the remaining weight was

distributed over the rational and emotional parts, ensuring that $part_{rat} + part_{emo} = 1$. Subsequently, we defined that

$w_{rat} = (1-w_{mor}) + part_{rat} * w_{rat\_opt}$ and
$w_{emo} = (1-w_{mor}) + part_{emo} * w_{emo\_opt}$,

where w_{rat_opt} and w_{emo_opt} represent the optimal weights for the rational and affective factors in making a decision.

3 RESULTS AND DISCUSSION

In a series of six simulations of actual medical cases, Moral Coppélia reached the same conclusions as expert ethicists [18]. A comment increasingly heard is that rational and logic approaches to moral decision making do not account for the whole plethora of moral choices that humans actually make. They are prescriptive models for formal ethical committees rather than descriptive of human behavior. Therefore, Pontier, Widdershoven, and Hoorn [20] analyzed the effects of affective factors on straightforward moral reasoning by letting Moral Coppélia function within the larger context of the affect model Silicon Coppélia. In employing the Trolley and Footbridge dilemma as our test cases, the influence of affect on moral reasoning could explain why moral dilemmas with a personal hue (i.e. Footbridge) lead to different decisions than more neutral cases (i.e. Trolley), although logically they are identical.

Within the ethical module of Silicon Coppélia as described by Moral Coppélia, autonomy (i.e. of the patient) appeared to be the top priority in moral verdicts. In zeroing in on autonomy as an ethical issue of its own, we distinguished between autonomy as self-determination and being free from the will of others (i.e. negative autonomy) and as the capacity of making a meaningful choice for one's own life (i.e. positive autonomy) [6]. We applied our twofold model of autonomy to legal cases of medical courts in The Netherlands [20]. Simulation results showed that the decisions of Moral Coppélia were congruent with the verdicts of health judges. Long-term positive autonomy sometimes was seen as more important than negative autonomy on the short-term, in which cases judicial coercion seemed to be justified.

Not only emotional aspects interfere with clear-cut rational choice in ethical dilemmas. Personality also can deflect from straightforward reasoning (e.g., [8]). This is why we borrowed the HEXACO model from criminology (e.g., [25]), in which (impaired) honesty and humility is one of the central predictors of criminal conduct.

Of course, verifying a system through simulation and finding comparable answers to real-life cases confirms – perhaps even proofs – the solidity of the logics and processes employed. It does not, however, make evident that there is significant ecological, read empirical, value or meaning to it. In a final study, then, we asked 153 participants to estimate the probability of making a criminal choice in four scenarios [19]. They assessed the perceived risk and negative state affect of decisions in a criminal situation. Also the participants' personality dimension of Honesty-Humility was measured. Then we connected an affective moral agent to each participant and tuned the parameters such that they optimally fitted the first half of the sample. With these parameter settings we could predict the criminal choice of the participants in the second half, the holdout sample, which succeeded pretty well. Prediction error turned out to be relatively low. The best predictions were achieved with the full model as compared to restricted models that either used the moral, rational, or affective factors. These findings correspond with current informal models of criminal decision making (i.e. [25]). Thus, we may enjoy some empirical evidence now that making a moral choice is dependent on personality, ratio, as well as feelings and that through affective Moral Coppélia with twofold autonomy and some personality aspects, we can simulate the process leading to such choices fairly well.

4 PRACTICAL CONSEQUENCES

"I am programmed to understand humans" is how android C-3PO reassures us in *Star Wars Episode II: Attack of the Clones*. And that is a very honorable cause for a programmer because currently, we are under attack of computer systems that run our lives autonomously – in pursuit of profit maximization, rationally, ruthlessly. From our understanding of humans, we contributed to the field of Machine Ethics by creating a moral robot that can take perspectives, switching on or off affective, personality, and rational aspects of moral decision making. Moral Coppélia with all her recently developed add-ons is the moral stop button that can be implemented into future personal assistants such as C-3PO so that our twofold autonomy is safeguarded, free from the robot's will, free to make a meaningful choice of how to employ our machine friends: In healthcare, for example, or criminal law.

If applied well, robots can make healthcare faster, better, cheaper, and more reliable not because they are but because they keep tedious or low-skill tasks away from the professional, who can then put her effort in quality, efficiency, and effectiveness. Or more importantly: In making healthcare more humane again. Reliability is not only a result of proper electro-mechanical functioning but also of ethical conduct. If grandma takes her cuddle bot to bed, we should make sure it does not bite when squeezed. Or more seriously, that it does not life stream its camera-eye recordings to Facebook. That it helps youngsters not to install illegal software on the robot's or any other device's hard drive – without acting like a pedantic principlist.

Like this, we can be sure to let our robots treat autistic children independently; have them positively impact mental syndromes, or reduce people's loneliness. With the latest Moral Coppélia installed, humanoid care robots or Caredroids uphold the dignity of patients and promote human values such as *caritas et iustitia*. As a partner in crime, moral robots may keep potential perpetrators from offence, because they are not only morally just; they are likeable; because they are your friend.

Acknowledgements. This study is part of the SELEMCA project within CRISP (grant number: NWO 646.000.003). We would like to thank Aimee van Wynsberghe and Joel Anderson for interesting discussions on earlier drafts of this paper.

REFERENCES

[1] Allen, C. Varner, G. & Zinser, J. 'Prolegomena to Any Future Artificial Moral Agent.' *Journal of Experimental and Theoretical Artificial Intelligence*, 12, 251–61 (2000)
[2] Anderson, M.; Anderson, S.; & Armen, C. 'MedEthEx: A Prototype Medical Ethics Advisor.' *Proceedings 18th Conference on Innovative Applications of AI*. Menlo Park, CA: AAAI Press (2006).
[3] Anderson, M., & Anderson, S. 'Machine ethics: Creating an ethical intelligent agent', *AI Magazine*, 28(4), 15-26 (2007)
[4] Banks, M.R., Willoughby, L.M., and Banks, W.A. 'Animal-Assisted Therapy and Loneliness in Nursing Homes: Use of Robotic versus Living Dogs.' *J American Med. Directors Assoc*, 9, 173-177 (2008)
[5] Beauchamp, T.L., Childress, J.F. 'Principles of Biomedical Ethics.' New York, Oxford: Oxford University Press (2001)
[6] Berlin, I. 'Two concepts of liberty.' Oxford: Clarendon Press (1958)
[7] Bosse, T., Pontier, M.A., Treur, J. 'A Computational Model based on Gross' Emotion Regulation Theory.' *Cognitive Systems Research Journal*, 11, 211-230 (2010)
[8] De Vries, R.E., Van Kampen, D. 'The HEXACO and 5DPT Models of Personality: A Comparison and their Relationships with Psychopathy, Egoism, Pretentiousness, Immorality, and Machiavellianism.' *Journal Personality Disorders*, 24, 244-57 (2010)
[9] Entwistle, V.H., Carter, S.M., Cribb, A., McCaffery, K. *J Gen Intern Med*, 25(7), 741–745. (2010)
[10] Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., Cohen, J.D. 'An fMRI Investigation of Emotional Engagement in Moral Judgment'. *Science*. 293, 5537, 2105-2108. (2001)
[11] Gross, J.J. 'Emotion Regulation in Adulthood: Timing is Everything' *Current directions psych. science*, 10(6), 214-219 (2001)
[12] Guarini, M. 'Particularism and Classification and Reclassification of Moral Cases.' *IEEE Intelligent Systems*, 21(4), 22–28. (2006)
[13] Haidt, J. 'The emotional dog and its rational tail: A social intuitionist approach to moral judgment.' *Psychological Review*. 108(4), 814-834 (2001)
[14] Hoorn, J.F., Pontier, M.A., Siddiqui, G.F. 'When the user is instrument to robot goals.' *Proceedings of 7th IEEE/WIC/ACM Conference on Intelligent Agent Technology*, pp. 296-301 (2008)
[15] Hoorn, J.F., Pontier, M.A., Siddiqui, G.F. 'Coppélius' Concoction: Similarity and Complementarity Among Three Affect-related Agent Models.' *Cognitive Systems Research Journal*, 33-49 (2012)
[16] Ohnsorge, K., Widdershoven, G.A.M. 'Monological vs Dialogical Consciousness – Two Epistemological Views on the Use of Theory in Clinical Ethical Practice.' *Bioethics*. 25, 7, 361-369 (2011)
[17] Picard R 'Affective computing.' MIT Press, Cambridge, MA (1997)
[18] Pontier, M.A., Hoorn, J.F. 'Toward machines that behave ethically better than humans.' *Proceedings of CogSci'12*, 2198-2203 (2012)
[19] Pontier, M.A., Van Gelder, J.L., De Vries, R.E., 'Computational Model of Affective Moral Decision Making that predicts Human Criminal Choices.' *Lecture Notes in Artificial Intelligence – PRIMA'13*, Vol. 8291, pp. 502-509, Springer Verlag (2013)
[20] Pontier, M.A., Widdershoven, G.A.M., Hoorn, J.F. 'Moral Coppélia - Combining Ratio with Affect in Ethical Reasoning.' *Advances in Artificial Intelligence – IBERAMIA 2012, Lecture Notes in Computer Science*, 7637, 442-451 (2012)
[21] Robins, B., Dautenhahn, K., Boekhorst, R.T., Billard, A.. 'Robotic Assistants in Therapy and Education of Children with Autism: Can a Small Humanoid Robot Help Encourage Social Interaction Skills?' *Journal of Universal Access in the Information Society*. 4, 105-120. (2005)
[22] Rzepka, R., Araki, K. 'What Could Statistics Do for Ethics? The Idea of a Common Sense Processing-Based Safety Valve.' *Machine Ethics: AAAI Fall Symposium*. Menlo Park, CA: AAAI Press (2005)
[23] Tronto, J. 'Moral Boundaries: a political argument for an ethic of care.' Routledge, New York. (1993)
[24] Van Gelder, J.L.: Beyond rational choice: The hot/cool perspective of criminal decision making. *Psychology, Crime & Law*. (2013)
[25] Van Gelder, J.L., De Vries, R.E. 'Traits and states: Integrating personality and affect into a model of criminal decision making.' *Criminology*, 50, 637-671 (2012)
[26] Van Gelder, J.L., De Vries, R.E., Van der Pligt, J. 'Evaluating a dual-process model of risk: affect and cognition as determinants of risky choice', *J Behavioral Decision Making*, 22, 45–61 (2009)
[27] Van Wynsberghe, A. 'Designing Robots for Care; Care Centered Value-Sensitive Design.' *Journal of Science and Engineering Ethics*, 19, 2, 407-433 (2013)
[28] Wada, K., and Shibata, T. 'Social Effects of Robot Therapy in a Care House', *JACIII*, 13, 386-392 (2009).
[29] Wallach, W., Allen, C., & Smit, I. 'Machine morality: Bottom-up and top-down approaches for modelling human moral faculties.' *AI and Society*, 22(4), 565–582 (2008)
[30] Wallach, W., Franklin, S. & Allen, C. 'A Conceptual and Computational Model of Moral Decision Making in human and Artificial Agents.' *Topics in Cognitive Science*, 2, 454–485 (2010)
[31] WHO 'Health topics: Ageing.' Available from: http://www.who.int/topics/ageing/en/ (2010)
[32] Widdershoven G.A.M., and Abma, T.A. 'Autonomy, dialogue, and practical rationality.' In: Radoilska, L. (ed.). *Autonomy and mental disorder*. Oxford: Oxford University Press, pp. 217-232 (2012).