AISB/IACAP World Congress 2012

Birmingham, UK, 2-6 July 2012

MORAL COGNITION & THEORY OF MIND

Marcello Guarini and Paul Bello (Editors)

Part of

Published by
The Society for the Study of
Artificial Intelligence and
Simulation of Behaviour

http://www.aisb.org.uk

ISBN 978-1-908187-22-2

Foreword from the Congress Chairs

For the Turing year 2012, AISB (The Society for the Study of Artificial Intelligence and Simulation of Behaviour) and IACAP (The International Association for Computing and Philosophy) merged their annual symposia/conferences to form the AISB/IACAP World Congress. The congress took place 2–6 July 2012 at the University of Birmingham, UK.

The Congress was inspired by a desire to honour Alan Turing, and by the broad and deep significance of Turing's work to AI, the philosophical ramifications of computing, and philosophy and computing more generally. The Congress was one of the events forming the Alan Turing Year.

The Congress consisted mainly of a number of collocated Symposia on specific research areas, together with six invited Plenary Talks. All papers other than the Plenaries were given within Symposia. This format is perfect for encouraging new dialogue and collaboration both within and between research areas.

This volume forms the proceedings of one of the component symposia. We are most grateful to the organizers of the Symposium for their hard work in creating it, attracting papers, doing the necessary reviewing, defining an exciting programme for the symposium, and compiling this volume. We also thank them for their flexibility and patience concerning the complex matter of fitting all the symposia and other events into the Congress week.

John Barnden (Computer Science, University of Birmingham)
Programme Co-Chair and AISB Vice-Chair
Anthony Beavers (University of Evansville, Indiana, USA)
Programme Co-Chair and IACAP President
Manfred Kerber (Computer Science, University of Birmingham)
Local Arrangements Chair

Foreword from the Symposium Chairs

The Society for Machines and Mentality – a special interest group of the International Association for Computing and Philosophy (IACAP) – is dedicated to advancing the philosophical understanding of issues involving artificial intelligence, philosophy, and cognitive science. This year's symposium was planned as something of a sequel to the symposium held at the 2011 IACAP meeting in Denmark. That meeting was dedicated to theory of mind, variously referred to as mindreading, mentalizing or mental state ascription. This year, we solicited papers devoted to moral cognition or theory of mind, as many believe there to be an important link between the two. Papers were submitted from 7 countries on four continents. Out of the 21 submissions made, 12 are scheduled for presentation, though not all are included herein. Authors who submitted their revised papers in the requested format have their papers included in these proceedings.

This symposium includes a scheduled debate; the speakers for that debate were invited, and they were not asked to submit papers, so their debate presentations are not included herein.

All papers in the proceedings for this symposium were refereed based on full paper submissions.

Marcello Guarini (Philosophy, University of Windsor, Canada)
Director of the Society for Machines & Mentality
Symposium Co-organizer and Co-Chair
Paul Bello (Office of Naval Research, USA)
Symposium Co-organizer and Co-Chair

Table of Contents

These papers are listed in the order in which they are scheduled to be presented at the conference.

Hard Problems: Framing the Chinese Room in which a Robot takes a Moral Turing Test

Wendell Wallach[1] **and Colin Allen**[2]

Abstract. Research on approaches for implementing moral decision-making capabilities within AI systems is contributing to a more comprehensive understanding of moral acumen. In addition to being able to reason, consciousness and understanding, a theory of mind, social skills, cooperating with other agents, the ability to solve frame problems, being embodied, empathy, and feeling pleasure and pain may be required for agents to successfully select morally acceptable courses of action within certain domains. The multifaceted nature of moral intelligence complicates both the task of designing artificial moral agents (AMAs) and the challenge of evaluating whether an artificial agent can make safe, legal, and appropriate decisions. Confronted with a somewhat comparable difficulty in determining whether a machine can think, Alan Turing [1] proposed his now famous imitation game. Fifty years later, Colin Allen, Gary Varner, and Jason Zinser [2] suggested a variant of the test: a moral Turing test (MTT). Within some limited domains, reason alone is sufficient for making moral judgments, and moral intelligence may be tested by using the traditional conversational method of posing a question and comparing the AI system's response to that of a human. However, evaluating an artificial agent's ability to accommodate a variety of moral considerations, including some that are difficult to quantify or describe, will necessitate testing it in complex situations. Consider the ability of an artificial agent to deduce the beliefs, desires, and intentions of other agents so that it might work cooperatively with them on a shared task, to distinguish a combatant from a non-combatant during guerrilla warfare, to discern that it is in a morally significant situation, or to discriminate which of many concurrent challenges it should respond to first. These tasks are not easy for humans, and yet we bring resources to bear in tackling such challenges that will be hard to reproduce artificially. Evaluating whether AI systems can manage such tasks sufficiently will either require special variants of the MTT, or they reveal the inadequacy of a MTT for evaluating the moral intelligence of artificial systems.

1 TWO HARD PROBLEMS

The task of building AI systems with even modest moral decision-making capabilities encompasses two hard problems. The first requires selecting a set of norms, rules, principles, or procedures for the system to use in making moral judgments, and finding a computational method to implement them. Most moral philosophers appreciate that finding an ethical[3] theory or rules to cover all cases adequately is itself a daunting, if not impossible, feat. Nevertheless, an analysis of the feasibility of instantiating an ethical theory or set of rules computationally is a step forward. Ethical theories, as well as procedures for facilitating the moral education of a machine, can be implemented through top-down and/or bottom-up approaches [3,2,4].[4] Presuming that one finds an adequate method to computationally implement an approach to ethical decision making, a further difficulty will be designing a method to test whether such a system will select safe, appropriate, and acceptable actions in real world situations.

The second hard problem concerns how to set boundaries to the assessments that must be carried out for effective moral decision making [7]. This is actually a group of related challenges. How does the system recognize that it is in an ethically significant situation? How does it discern essential from inessential information? How does the AMA estimate the sufficiency of initial information? What capabilities would an AMA require to make a valid judgment about a complex situation, e.g., combatants vs. non-combatants? How would the system recognize that it had applied all necessary considerations to the challenge at hand or completed its determination of the appropriate action to take? For example, what stopping procedure would the system use to determine that it had completed a utilitarian calculation?

To be sure, humans can sometimes fail in all of the ways implicit in these questions, for instance making mistakes in tasks such as determining who is a combatant and who is a non-combatant. Nevertheless, humans bring powers of discrimination to bear to performing such tasks that we either do not know how to implement in robots, or for which we have at best a few rudimentary theories.

This group of challenges is related to the frame problem, both as it was first elucidated by AI researchers [8] and as it was later embellished by philosophers [9,10] to cover wider epistemological issues. In AI, the frame problem concerns how to represent only those effects of an action that are relevant to choosing among actions without having to also explicitly represent all the intuitively obvious mundane effects. For philosophers the problem extended to how any intelligent agent would limit the set of beliefs that must be re-evaluated and possibly changed as the result of an action.

[1] Yale University, Interdisciplinary Center for Bioethics, New Haven, CT 06520-8293, USA. Email: wendell.wallach@yale.edu

[2] Dept. of History and Phil. of Science, Indiana University, Bloomington, IN 47405, USA. Email: colallen@indiana.edu

[3] In this paper the terms 'moral' and 'ethical', as well as their variants, will be used interchangeably.

[4] Bottom-up or developmental approaches take their inspiration from the very same paper in which Turing [1] first introduces the imitation game. Connectionist bottom-up approaches may lead to moral sensitivities that cannot be reduced to moral principles [5,4,6].

Frame problems arise in implementing any norms, rules, principles, or procedures in an AMA. An AMA functioning in anything other than a tightly bounded context will carry a heavy computational load as it will need to estimate the sufficiency of the initial information available and search out sources for additional information, it will be required to have significant psychological knowledge about the other actors in the environment, and it will need to have knowledge of effects of actions (its own and that of other actors) in the world. The difficulty for an AMA is that the boundaries for evaluation of its possible actions are potentially unlimited [11]. Nevertheless, humans manage to function with a number of heuristic and affective processes [4,11] that effectively limit the kinds of unlimited search that seem to be demanded by more formal procedures. How to implement these in AI is unclear, a point to which we return below.

Moral philosophers and applied ethicists tend to focus upon determining the norms, rules, principles, or procedures for making moral judgments. Generally, the cognitive mechanisms that serve frame problems are presumed as givens when discussing moral decision-making by human actors. However, research in the cognitive sciences has reinvigorated the study of moral psychology, focused renewed attention on the role of various unconscious mechanisms in facilitating moral judgments, and fostered a re-evaluation of the *is/ought* distinction [12]. The two hard problems certainly overlap but neither subsumes the other. For example, in all but the simplest contexts AMAs will need to solve frame problems in order to determine what information will need to be factored into decisions where rules, principles, or procedures are applied.

2 IMPLEMENTING SUPRARATIONAL MECHANISMS

Applying reason to factual information is useful and important but not sufficient for ensuring that humans will act in a morally intelligent manner. Expunging emotional influences from moral decisions was an ideal for the Stoic philosophers. However, we live in the age of emotional intelligence [13,14,15], cognizant that affect is intricately bound up with what passes for reasoning [16,17]. Emotions sensitize us to moral considerations and facilitate quick responses to challenges. They are also among the cognitive mechanisms that contribute to moral intelligence by facilitating responses to challenges where information is incomplete, unclear, or inaccurate.

Early research has made it clear that implementing moral decision making within AI system will in many domains need to be supported by not only emotions, but also consciousness, a theory of mind, being embodied in a world with other entities, and additional suprarational capabilities [4,18]. The actions taken by AMAs will be judged by human standards; however, AMAs will not necessarily emulate human capabilities in order to respond to morally significant situations in a safe and appropriate manner. Nevertheless, AMAs will need to be equipped with cognitive mechanisms that functionally emulate the various suprarational capabilities that serve moral decision making in humans. We can expect cross-fertilization between the study of these cognitive mechanisms in humans and the need for additional mechanisms to facilitate computational systems making acceptable moral judgments.

Suprarational mechanisms will be necessary for both the hard problem of implementing norms, rules, principles, or procedures and for the hard *frame problems*. The suprarational mechanisms support the two hard problems in a number of ways including: gathering information, recognizing the need for additional information, integrating sensory input, and focusing attention upon significant features of a complex situation. Computer scientists are actively theorizing about the design of suprarational cognitive mechanisms and preliminary experiments directed at implementing components of suprarational mechanisms are underway. However, the various theories are far from proven, the existing implementations are rudimentary, and the future task of integrating the subsystems that support a suprarational capability, such as the so-called "theory of mind" (ToM), are daunting.

2.1 Theory of Mind and Empathy

Let us consider the implementation of a ToM in greater depth. In humans, a suite of capacities relevant to ToM develops through the early years of life. An infant learns in stages to distinguish her own body from that of others, to recognize herself in a mirror (primitive self-awareness), and to appreciate that another's mind will contain different information from her own [19]. All of these contribute to the development of ToM, usually defined as the ability to attribute mental states (beliefs, intents, knowledge, desires, etc.) to oneself and others, and to appreciate that the content of other minds differs from one's own. A ToM is foundational for deducing the beliefs, desires, and intentions of other, engaging others socially in characteristically human ways, and cooperating in complex shared tasks.

The research on ToM is filled with fascinating experiments and characterized by an array of largely unproven theories. Nevertheless, AI engineers are already testing these theories in the design of their robots. The computer scientist Brian Scassellati was among the first to consider developing a ToM for a robot and wrote his PhD thesis at M.I.T. on the subject [20]. Now at Yale, Scassellati continues this work on ToM with the development of the robot Nico, designed to model a human infant.

ToM is often presumed to emerge from a collection of very low-level skills. For example, in 1944 Mary-Ann Simmel and Fritz Heider demonstrated with a few simple video clips, that people impute intentions to objects based on simple movements. Associating intentions with basic movements is one of the lower level skills that should also contribute to building a full ToM in an AI system. Identifying basic emotions with gestures and expressions on the face of other actors is another.

Utilizing the theories of cognitive scientists who have broken ToM down into discrete skills, computer scientists are trying to implement each of these skills in hardware and software. For example, humans distinguish sensory inputs that are a result of their own actions from sensory inputs that arise from the actions of others. Kevin Gold and Brian Scassellati [21] demonstrated how the timing of sensory feedback after self-generated movement can be used to enable their robot Nico to distinguish sensory inputs produced by its own actions from those produced by the movements of others.

Current research on building a robot with ToM is proceeding on the assumption that the aggregation of lower level cognitive mechanisms will collectively enable the robot to act as if it had a

ToM. To date, only a few of the basic skills that contribute to ToM have been instantiated computationally. Identifying the full skill set that contributes to ToM, and the hard work of coordinating or integrating these skills lies ahead. To date we not only lack AI systems with a ToM, but we do not even know whether we have adequate theories about the attributes that are necessary for a system to have a ToM. The jury is out on whether this reductionistic behavioural-based approach to building an artificial ToM will work. Nevertheless, the first steps taken by Scasselati and his colleagues are impressive enough to suggest that significant strides can be anticipated over the next decade.

ToM and empathy are related, but the relationship between these two concepts is far from clear. Certainly both contribute to the way one human appreciates the states of mind of another. The capacity to empathize with the feelings of others is often considered to be a prerequisite for moral judgment and sensitive behaviour in a variety of situations where people interact. Nevertheless, there are many cases of psychopaths who are skilled at deducing appropriate empathetic behaviour without actually feeling empathy [22,23].

Empathy would enhance an artificial system's ability to select morally appropriate responses in its choice of actions. However, robots are unlikely to have empathy for humans or non-human animals unless or until they have emotions of their own. Without emotions, empathic behaviour by robots will be largely the result of rational responses (cognitive emotions) built on top of a merely symbolic representation of the minds of others. Presumably deductions about the emotions of others by AMAs would be directed towards more praiseworthy goals than those of their psychopathic counterparts. Whether robots will require a full somatic architecture or even need to be biological organisms to be truly empathetic is an outstanding question. Perhaps cognitive emotions will be sufficient for AI systems to behave morally, but whether this is truly the case can only be known after building and testing AMAs.

3 A MORAL TURING TEST

Before artificial systems can be deployed, an evaluation needs to be made as to whether that entity can perform safely and appropriately within the context where it will work. The artificial agents deployed to date have operational morality [4]. The designers and engineers predict the situations the system will encounter and outfit it with sensors and software designed to facilitate the robot's ability to respond safely and appropriately to the anticipated challenges. However, increasingly autonomous agents will need to be explicit moral agents [24] or functionally moral [4]. Systems that are autonomous in the sense that they act without direct human intervention will need to be able to evaluate various courses of action and select the best response to a challenge on their own.

Testing the moral intelligence of operationally moral agents that maneuver within very constrained contexts is relatively straightforward. However, testing the moral intelligence of increasingly autonomous artificial agents will be difficult, and perhaps impossible, particularly as AMAs approach being considered full moral agents with rights and responsibilities.

Sidestepping the problem of defining intelligence, Alan Turing [1] proposed an imitation game, better known as a Turing test. Turing himself responded to criticisms of the test he anticipated in the original paper, and many additional critiques have been made over the past sixty years. Does a machine have to imitate a human in order to think? Would a machine have to dumb itself down in certain domains in order to make it more difficult for the expert to distinguish the machine from the human? Perhaps the most serious critique, or at least the one which has received the most attention, is that of John Searle [25], who argued that a machine might succeed at the imitation game while, for example, translating Chinese, without any *understanding* of the task it was performing. Searle's Chinese Room example was meant to illustrate that the manipulation of symbols by computers should not be confused with semantic understanding. The Chinese Room has received its own share of criticism, but has, nevertheless, served to stimulate serious philosophical refection on what is meant by 'knowledge' and 'understanding.'

There was considerable concern about the value of the Turing test when in 1966 ELIZA, a program created by Joseph Weizenbaum [26], fooled some people into believing that it was human. However, the test has endured. The Turing test continues to play an important philosophical role in AI [27,28] even if its use for evaluating whether a machine can think is far from perfect. No one has suggested a better test.

Fifty years after Turing first proposed the imitation game, Allen, Varner, and Zinser [2] proposed a Moral Turing test (MTT) and made some critical observations regarding its feasibility. A key advantage of an MTT is that it would help bypass ethical disagreements. We commonly hold that other humans are moral agents even when their values differ from ours, presuming that they can offer acceptable justifications for their actions.

However, there is likely to be resistance to deploying an AMA that is only as good in making moral judgments as a fallible human. Exceeding human moral judgment, and thereby being distinguishable from a human, would be a plus. Thus Allen, Varner, and Zinser propose a comparative MTT (cMTT for short) in which the interrogator judges "Which of these agents is less moral than the other?" Of course a cMTT would also set too low a standard for the machine to be judged a moral agent if the human respondent was less than a paragon of virtue. No psychopaths need apply for the role of human respondent. A cMTT might suffice if the human respondent is certified to be a person of high moral character and sensitive to a broad array of moral considerations.

4 DEMONSTRATING MORAL APTITUDE IN COMPLEX SITUATIONS

Conversational and analytic intelligence may be adequate for moral decision making in constrained contexts where all the necessary information is readily available. It may also be acceptable for certain kinds of decisions in business or in public policy discussions where utilitarian calculations are applied to optimizing a predefined measure of utility.

But moral intelligence as it is applied to a broad array of complex challenges is something more than conversational and analytic intelligence. Given the multifaceted and multi-dimensional nature of moral decision making, an adequate MTT would require comparing agents responding to real-time challenges while being embodied and situated in a rich

environment. In the following sections we will use four examples that illustrate both the complications of designing and engineering AMAs and the difficulty of designing an MTT that tests specific aspects of moral intelligence.

4.1 Cooperation

The ability to deduce the knowledge, beliefs, desire, and intentions of the other people you are working with is central for success in completing a shared task. It is also important when humans and machines work on a task together.

No robot is an island. Bots in computer networks and robots will be embedded in a sociotechnical system [29]. AMAs cannot be designed properly without attention to the systems in which they are embedded. Rather than designing more sophisticated capacities for the robots themselves, sometimes the better approach is to rethink the entire edifice that produces and uses them. Deborah Johnson has patiently and persistently insisted at various conferences and workshops that focus on the capabilities of the robots considered as independent artefacts carries potential dangers, insofar as it restricts attention to one kind of technological fix instead of causing reassessment of the entire socio-technological system in which bots and robots operate.

In a similar vein, David Woods and Erik Hollnagel [30,31] maintain that with the exception of a few limited purpose machines, an intelligent system and the operators who work with it are best understood as a Joint Cognitive System (JCS). JCSs require tight coordination between the activities of the human and the mechanical components. Given that the actions of the mechanical components tend to be limited, there is usually added reliance on the flexibility of the human operators.

Woods and Hollnagel [30] note that with the advent of artificial agents, when a JCS fails there is a tendency to blame the human as the weak link, and to propose increased autonomy for the mechanical device as a solution. Furthermore, there is the illusion that increasing autonomy will allow the designers to escape responsibility for the actions of artificial agents. However, Woods and Hollnagel point out that increasing autonomy will actually add to the burden of human operators. They illustrate this with the example of an accident on December 6th, 1999, that caused $5.3 million in damages when there was a failure in coordination between operators and a semi-autonomous Global Hawk UAV. Manoeuvring the Global Hawk on the ground, the operators misunderstood the system's actions. The conflict between what the system was doing and what the operators thought the system was doing led to the aircraft going off the runway, where its nose gear collapsed. The focus on isolated machine autonomy distorts the full appreciation for the kinds of systems design problems inherent in JCSs.

The behaviour of robots will continue to be brittle on the margins as they encounter new or surprising challenges. Human operators will need to anticipate what the robot will try to do in new situations in order to effectively coordinate their actions with those of the robot. However, anticipating the robot's actions will be harder to do as systems become more complex and independent, leading to a potential increase in conflicts between the actions initiated by the system and the actions initiated by the human operators. While each failure may be attributed to the operators, to expect operators to anticipate the actions of intelligent systems becomes more and more unreasonable as the systems and the environments in which they operate become more complex.

The central point we wish to underscore here, is that evaluating the moral aptitude of a machine will in many domains have more to do with understanding how it functions as a component in a sociotechnical system and as a member of a team than as an autonomous entity making moral decisions. If the goal is the safety and success of the enterprise rather than the special intelligence of the AI system, than testing for cooperative skills can often tell us more than testing for independent reasoning abilities. This would suggest versions of MTT in which a team leader makes a decision about which of two entities, an intelligent system or a human, she would like to have on her team.

For a partnership or team to work, all members need to be able to deduce the goals, beliefs, and intentions of each other. Accidents like the Global Hawk collapse could be reduced if the system also had a way of understanding what the operators were trying to do, and accommodating their intentions in its actions. Over the next decade or two, this kind of coordination in which man and machine function in a close partnership is likely to be a particularly fruitful approach for engineering systems that function in a safe and moral manner.

4.2 Distinguishing Friend From Foe

The adoption of drones and unmanned ground robots, such as IRobot's PackBot, by military forces around the globe has placed a focus on the ethical use of AI during warfare. Ronald Arkin, contends that robots can be developed that will follow the Laws of War and the Terms of Engagement and furthermore will be more ethical than human soldiers [32]. Arkin acknowledges that he is only tackling the first hard problem discussed above, and therefore, the ethical governor he is attempting to design will only facilitate robots behaving morally within tightly constrained military contexts, not while fighting guerrillas or revolutionaries in urban landscapes.

The dangers inherent in autonomous systems initiating kill orders are a central concern for critics of military robots [33-39]. They commonly point out the fact that present day robots lack situational awareness and are unable to distinguish combatants from non-combatants. Nor will robots be likely to have the necessary capabilities to perform these tasks in the near future. Distinguishing friend from foe, for example, is also a difficult challenge for humans, but we bring cognitive resources to bear on the problem that are unavailable to robots.

Police forces, such as the FBI, and the military have developed real-world simulations and simulators of virtual worlds that help train personnel in distinguishing friend from foe. Presuming that robots will improve in this ability, simulations and simulators might be used to test their success and compare it to that of their human counterparts. Interestingly, humans might perform better within simulated environments than they will in the real world. The knowledge that incorrect actions within a simulation will not lead to harm or death to other humans or themselves dampens some of the intense emotions associated with being in actual combat. On the other hand, the behavior in a simulator by a robot that lacks somatic emotions may match fairly closely to its behavior outside of the simulation.

The point of this example is not that it would be difficult to test whether a robot is adequately skilled in the task. Rather, we wish to point out here:

1) This test could not be performed using the normal question and answer format.

2) In measuring acuity in ethically significant tasks the testing environment will only moderately match the real world and this will alter the behaviour of humans and/or robots. It is of course their behaviour in the real world that should be our greatest concern.

4.3 Discerning a Moral Challenge

How will a robot discern that it is in a morally significant situation [40]? Asking an AMA to answer to trolley problems or other philosophically inspired thought experiments is a far cry from placing a system in a rich social environment and observing whether it recognizes and responds appropriately to a morally significant situation. Consider walking down the street in a busy city when an older man steps off the curb heading into moving traffic or a youngster grabs a women's purse and begins to run away. Many people would recognize these scenarios as morally significant situations that would require a response. At the least, a sensitive moral actor would make a quick judgment as to whether another actor is responding, or whether to call for help in order to elicit a response from a police officer or some other actor. Most people would reach out to pull the older man back to the curb, and some might even chase down the thief. Presuming that an AMA actually discerns the moral challenge occurring in the midst of the bustling thoroughfare, it too might well act appropriately. But would the AMA pick out the morally significant scenario? How would it discern the salient from the insignificant information? What cognitive architecture would be required to alert the AMA that this is an event that requires its attention and action?

How would we evaluate whether the AMA had the intelligence to discern a morally significant event within a rich social environment? Perhaps this could be tested for using a virtual reality simulation. But human agents, with whom the AMA will be compared, come equipped with affective activation mechanisms that are likely to function differently in real world situations than they would in a simulation. Even if we develop affective mechanisms for machines they are unlikely to function in the same way as human emotions. Finally, experimental attempts to place humans and robots in similar situations could be inadequate because it will be difficult, if not impossible, to reproduce exactly the same situation and set of influences for both respondents.

4.4 Prioritizing Moral Challenges

Procedures for prioritizing moral considerations have been developed for specific contexts such as triage during admission to a Hospital Emergency Room or for EMTs responding at the site of a disaster. Presumably an AMA could learn these procedures and also be programmed with accompanying expert systems that would facilitate a high level of performance. However, the challenge is more difficult when we consider free-roaming artificial agents.

As increasingly autonomous AMAs navigate through complex social environments they will encounter countless ethical challenges. Arguably ethical considerations arise for a very broad range of tasks, for example, when values are used: 1) to fill in for information that is incomplete, unclear, or inaccurate 2) for prioritizing the relative importance of the information available. However, we will not want an AMA to become absorbed by trivial endeavors and will rather want it to prioritize and work on the more essential tasks. Protecting others from immediate harms or self-preservation should certainly take precedence over searching for a missing piece of information required for a future task.

How do humans prioritize the array of challenges that come to their attention? As it encounters a sea of information, how will the AMA pick out those concerns that are most worthy of its attention? The neuroscientist Bernie Baars [41,42] proposed a theoretical answer he named Global Workspace Theory (GWT) to the human side of this equation. GWT outlines a functional role for consciousness in picking out which of many competing cognitive inputs wins the battle for consciousness and thereby gets our attention.

GWT has caught the attention of a number of computer scientists including Stan Franklin [43], who together with Baars and other colleagues has formulated a conceptual and computational model of GWT called LIDA. LIDA models how an agent makes sense of its world and figures out what to do next. Together with Wallach and Allen, Franklin [4, 44, 45] has explored adapting LIDA to discern, prioritize, and solve moral challenges. One outstanding question is whether the functional role for consciousness instantiated in a LIDA-based AMA would be adequate for the system to solve moral challenges, or whether the AMA would also require the phenomenal attributes of consciousness [46, 45].

LIDA is merely one example of how a system with artificial general intelligence (AGI) might solve the problem of prioritizing vast quantities of information and selecting the tasks to focus upon. Other AGI systems might prioritize in a different manner. What was important to Wallach, Franklin, and Allen was to at least offer a conceptual way forward. That way forward would require an AGI that is embodied and embedded in a rich environment. Testing such a system will also require a similarly rich environment, not a sterile room where an isolated computer engages in exchanging messages with an interrogator.

5 CONCLUSIONS

Many technological thresholds will need to be crossed before we are able to build artificial systems capable of functioning autonomously within several domains. Full moral agency for a computational system is an even more distant prospect, as will be the need for a general purpose MTT. In the meantime, systems will be developed for morally bounded contexts and the testing of their moral intelligence will be restricted to functioning safely and appropriately within those contexts. Variants of the MTT might be designed for testing the moral intelligence of AMAs within these restricted domains.

However, the traditional conversational variants of the Turing test will not be adequate for evaluating many dimensions of moral acumen. Moral intelligence underscores the significant role played by suprarational capabilities for ensuring that an agent is sensitive to a wide range of moral considerations.

Testing whether AMAs have these moral sensitivities will require special tests within rich social environments and may even preclude certain one-on-one comparisons with human agents. Does this mean it will be impossible to design an environment, a proverbial "room", in which the MTT is conducted? Not necessarily. It is just that we need to keep in mind that systems passing comparative tests in more bounded contexts have only provisionally demonstrated their moral intelligence.

ACKNOWLEDGEMENTS

Thanks are owed to the referees for helpful comments that certainly contributed towards improvements in this paper.

REFERENCES

[1] A. Turing. Computing Machinery and Intelligence. *Mind*, 59 (236): 433–60 (1950).

[2] C. Allen, G. Varner and J. Zinser. Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical Artificial Intelligence*, 12, 251–261 (2000).

[3] J. Gips. Towards the Ethical Robot. In: *Android Epistemology*. K.G. Ford, C. Glymour, and P.J. Hayes (Eds.). Cambridge, MA: MIT Press (1991).

[4] W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right From Wrong.* New York: Oxford University Press (2009).

[5] J. Dancy. *Ethics Without Principles.* Oxford: Clarendon Press (2004).

[6] M. Guarini. Computational Neural Modeling and the Philosophy of Ethics. In: *Machine Ethics*. M. Anderson and S. Anderson (Eds.). Cambridge: Cambridge University Press (2011).

[7] W.Wallach and C. Allen. Framing Robot Arms Control. (submitted).

[8] J. McCarthy and P.J. Hayes. Some Philosophical Problems from the Standpoint of Artificial Intelligence. In: *Machine Intelligence 4.* D. Michie and B. Meltzer (Eds.). Edinburgh University Press (1969).

[9] D. Dennett. *Brainstorms.* MIT Press.(1978).

[10] J.A. Fodor. *The Modularity of Mind*. MIT Press (1983).

[11] M.Guarini and P. Bello. Robotic Warfare: Some Challenges in Moving from Noncivilian to Civilian Theaters. In: *Robot Ethics: The Ethical and Social Implications of Robotics*. P. Lin, K. Abney and G.A. Bekey (Eds,). MIT Press (2012).

[12] W. Sinnott-Armstrong (Ed.). *Moral Psychology* (3 volumes). MIT Press (2007, 2008).

[13] A. Gardner. *Frames of Mind: The Theory of Multiple Intelligences.* New York: Basic Books (1983).

[14] P. Salovey and J.D. Mayer. Emotional Intelligence. *Imagination, Cognition and Personality,* Vol. 9(3), 185-211 (1990).

[15] D. Goleman. *Emotional Intelligence: Why It Can Matter More Than IQ.* Bantam Books (1996).

[16] A. Damasio. *Descartes Error: Emotion, Reason and the Human Brain.* New York: Putnam (1994).

[17] J. Panksepp. *Affective Neuroscience: The Foundations of Human and Animal Emotions.* Oxford: Oxford University Press (1998).

[18] W. Wallach. Robot Minds and Human Ethics: The need for a comprehensive model of moral decision making. *Ethics and Information Technology,* 12: 243-250 (2010).

[19] H. Wimmer and J. Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13,* 103-128 (1983).

[20] B. Scassellati. Foundations of a theory of mind for a humanoid robot. Ph.D. dissertation, MIT Department of Computer Science and Electrical Engineering (2001).

[21] K. Gold and B. Scassellati. Using Probabilistic Reasoning Over Time to Self-Recognize. *Robotics and Autonomous Systems* (RAS) 57(4), p. 384-392 (2009).

[22] H.M. Cleckley. *The Mask of Sanity: An Attempt to Reinterpret the So-Called Psychopathic Personality*, 5th Edition. ([1941] 1984).

[23] R.D. Hare. *Without Conscience: The Disturbing World of the Psychopaths Among Us*. New York: Guilford Press (1999).

[24] J.H. Moor. The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems* 21(4), 18-2 1 (2006).

[25] J. Searle. Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), 417–57 (1981)

[26] J. Weisenbaum. ELIZA-A Computer Program for the Study of Natural Language Communication Between Men and Machines. *Communications of the ACM*, 9: 36–45 (1966).

[27] A. Saygin, I. Cicekli and V. Akman. Turing Test: 50 Years Later. *Minds and Machines*, 10: 463–518 (2000).

[28] J.H. Moor. The Status and Future of the Turing Test. *Minds and Machines*, 11: 77–93 (2001).

[29] D. Johnson and K.W. Miller. *Computer Ethics: Analyzing Information Technology, 4th Edition.* Prentice Hall (2009).

[30] D.D. Woods and E. Hollnagel. *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering*. CRC Press (2006).

[31] E. Hollnagel, D.D. Woods and N. Leveson (Eds.). *Resilience Engineering: Concepts and Precepts.* Ashgate Publishing (2006).

[32] R. Arkin. *Governing Lethal Behavior in Autonomous Robots*. Chapman and Hall/CRC (2009).

[33] P. Asaro. How Just Could a Robot War Be? In: *Current Issues in Computing And Philosophy.* P. Brey, A. Briggle and K. Waelbers (Eds.). Amsterdam, The Netherlands: IOS Press (2008).

[34] J. Borenstein. The Ethics of Autonomous Military Robots. *Studies in Ethics, Law, and Technology*, 2 (1): Article 2. DOI: 10.2202/1941-6008.1036 (2008).

[35] J. Altmann. Preventive Arms Control for Uninhabited Military Vehicles. In: *Ethics for Robotics*. R. Capurro and M. Nagenborg (Eds.). AKA Verlag, Heidelberg (2009).

[36] A. Krishnan *Killer Robots: Legality and Ethicality of Autonomous Weapons.* Burlington: Ashgate (2009).

[37] R. Sparrow. Predators or Plowshares? Arms Control of Robotic Weapons. *IEEE Technology and Society* 28 (1): 25-29 (2009).

[38] N. Sharkey. The Automation and Proliferation of Military Drones and the Protection of Civilians. *Law, Innovation and Technology,* (3)2, pp. 229-240 (2011).

[39] N. Sharkey. Killing Made Easy: From Joysticks to Politics. In: *Robot Ethics.* P. Lin, K. Abney, and G.A. Bekey (Eds.). MIT Press.(2012).

[40] D. McDermott. Why Ethics is a High Hurdle for AI. Paper presented at the 2008 North American Conference on Computing and Philosophy, May 12, Bloomington, Indiana (2008).

[41] B. Baars. *A Cognitive Theory of Consciousness.* Cambridge, MA: Cambridge University Press (1988).

[42] B. Baars. *In the Theater of Consciousness.* New York, NY: Oxford University Press (1997).

[43] S. Franklin. A "Consciousness" Based Architecture for a Functioning Mind. In: *Visions of Mind*, D. Davis (Ed.). Hersey, PA: Idea Group, Inc. (2001).

[44] W. Wallach, S. Franklin and C. Allen. A conceptual and computational model of moral decision making in human and artificial agents. *TopiCS in Cognitive Science,* Vol 2 (3), 454-485 (2010).

[45] W. Wallach, C. Allen and S. Franklin. Consciousness and Ethics: Artificially Conscious Moral Agents. *International Journal of Machine Consciousness*, 3(1), 177-192 (2011).

[46] S. Torrance. Ethics and consciousness in artificial agents. *Artificial Intelligence and Society*, 22(4), 34 (2008).

Moral action and mechanical models of intelligence: What can we learn from the Turing Test?

J. A. Quilici-Gonzalez[1], M. C. Broens[2], G. Kobayashi[1] and M. E. Q. Gonzalez[2]

Abstract. In the present paper, two main questions are addressed: 1) What can we learn about moral actions from the Turing Test of intelligence? 2) What could be the advantages, disadvantages, and difficulties of implementing a Moral Turing Test in virtual systems that include disguisers (these are programs capable of creating artificial virtual entities that may ostensibly present aspects of real persons to humans with whom it interacts)? In the context of human-disguiser communication, we argue that deontological and utilitarian ethics cannot fully accommodate the kind of problems present in contemporary Ethics, which requires a more flexible and fuzzy approach to morality. We focus also on the distinction between habits and intelligent moral capacities in order to evaluate problems related to the implementation of a moral Turing Test in virtual systems. Based on the notion of the *captcha* test (Completely Automated Public Turing Test to tell Computers and Humans Apart), we inquire into the possibility of adjusting this test to identify fuzzy elements that could characterize responses of hybrid human-disguiser systems.

1 INTRODUCTION

Moral actions seem to incorporate acquired dispositions involving training, critical attention and creativity from an agent situated in a social and environmental system. From this supposition, two main questions are addressed in the present paper: (a) what can we learn about moral actions from the Turing Test of intelligence [1], and (b) what could be the advantages, disadvantages, and difficulties of modeling a Moral Turing Test (MTT) in virtual systems that include disguisers? As illustrated in Figure 1, disguisers are programs capable of creating artificial virtual entities that may ostensibly present aspects of real persons to humans with whom they interact [2].

In the context of human-disguiser communication, taking into consideration that the Turing Test is based on language, and not on action, or on descriptions of actions, we suggest that disguisers could minimize this restriction by allowing the introduction of sounds, images, and movement in the communication. From this mixed perspective, we stress the limitations of virtue, deontological and utilitarian ethics in dealing with contemporary moral action in the domain of human-machine interaction.

[1] Federal University of ABC, Rua Santa Adelia 166, Santo Andre, SP, 09210-170 Brazil. Email: jose.gonzalez@ufabc.edu.br, guiou.kobayashi@ufabc.edu.br.

[2] University of Sao Paulo State, Av. Hygino Muzzi Filho 737, Marilia, SP, 17525-900, Brazil. Email: mariana.broens@gmail.com, gonzalezmeq@yahoo.com.br.

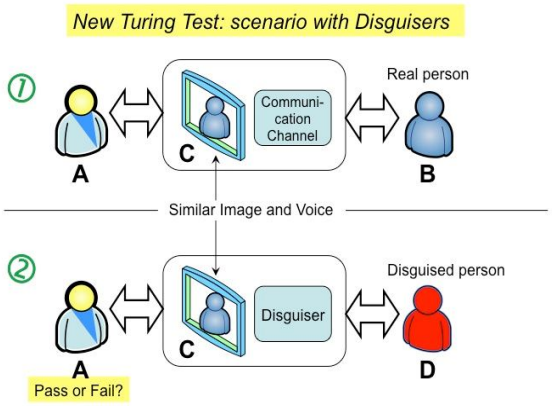

Figure 1. New Turing Test with Disguisers.

In order to analyze the question concerning the advantages, disadvantages, and difficulties of modeling a Moral Turing Test in virtual systems that include disguisers, we focus on the concept of affordance [3] and the *captcha* test (Completely Automated Public Turing Test to Tell Computers and Humans Apart). The *captcha*test is used in computing as an attempt to guarantee that a response to a question is generated by a person and not by a machine, and here we inquire into the possibility of adjusting this test to identify invariant elements that could characterize the moral responses of hybrid human-disguiser systems.

2 MORAL ACTIONS AND THE TURING TEST

The Turing Test, as originally described in the 1950 paper, is founded on language; it is by means of a set of descriptions of intelligent behavior that a machine might be able to demonstrate its ability to think in a way indistinguishable from that of an intelligent human being. As the language basis of the Turing Test imposes strong restrictions on the scope of the investigation of moral actions, disguisers could improve the test, using facial expressions, body movements, and voice intonations. With these new resources, the disguiser's main role would lie in describing, or representing in a theatrical fashion, a sequence of human actions. It would be the task of the interrogator to ask the disguiser technical questions (such as details of the circumstances in which an event occurred), and then separate the

actions of the machine from those of the human behind the disguiser.

However, by adding new elements to contemporary human-disguiser communication, new ethical problems arise, related to the development of the Turing Test of intelligence, which require answers that seem to go beyond those provided by traditional Ethics. In order to understand this point, we introduce some premises of virtue, deontological and utilitarian ethics, as a way to investigate these new problems.

According to Aristotelian virtue ethics [4], the "character" (or the set of personal dispositions), molded by customary patterns of action, is the main element of virtuous action: each person acts according to their own virtues. In the collective, social, perspective there is a general tendency to goodness, thanks to the improvement in morality of agents learning by example. This tendency is a result of the long-term dynamics of causal relationships between entities, according to the Aristotelian theory of four causes (material, formal, efficient, and final). As pointed out by Gonzalez et al. [5, p. 377], the material cause corresponds to the basic elements constitutive of organic and inorganic matter; the formal one concerns their structure and function; the efficient cause sets the changes and all forms of motion, and the final cause is related to the direction in which changes occur.

We believe that an Aristotelian-inspired perspective of ethics and causality would not be fully suitable for investigation of the social relationships mediated by information technology (such as the disguisers themselves), because the form-matter separation, common in the virtual world, may disrupt the causal dynamics that makes possible the practical learning of virtue. In other words, to learn and to perform a virtuous action, the agent needs some contextual and pragmatic components that the virtual relationship does not seem to incorporate. In this sense, Aristotelian virtue ethics would be inappropriate to deal with ethical issues related with virtual human relationships mediated by disguisers.

In turn, as indicated by [6], according to Kantian deontological ethics, every human action must be guided by respect for rules and duties, in a mandatory fashion. Hence, an action will only be considered morally acceptable if it can be transformed into a universal law, a categorical imperative that is valid in every similar situation, independent of the immediate consequences that may follow from it. Kant's famous categorical imperative: "Act only in accordance with that maxim through which you can, at the same time, will that it become an universal law" [7] characterizes this principle, as illustrated in the following example: One is not allowed to lie in any situation whatsoever, because the generalization of such practice would render unfeasible the creation or survival of a rational society with moral laws.

In the context of disguiser-human communication, the acquisition of any false virtual identity would not be considered morally acceptable, according to deontological ethics, because it could represent a form of lie; a disguised person would be hiding his/her real identity, using an unaware receiver as a means to accomplish an end that could eventually bring benefits only to the sender. However, this kind of disguised communication is becoming increasingly common nowadays, and it seems problematic to consider it as simply immoral.

As an alternative to deontological ethics, the utilitarian ethics proposed by Jeremy Bentham [8] states that, every human being should consider how the consequences of his/her actions could increase the common good or happiness in society. Instead of a universal moral law, utilitarian ethics suggests flexible rules adapted to each case, provided that common happiness is taken into consideration. Disguiser-human communication might therefore be morally acceptable if it could produce benefits for the community. However, what criteria of relevance should be elected to evaluate these benefits? We are faced here with the difficult problem of consensus on the criteria of relevance that should guide and delimit the scope of moral benefits.

As stressed by Beavers [9]:

> (…) the project of designing moral machines is complicated by the fact that even after more than two millennia of moral inquiry, there is still no consensus on how to determine moral right and wrong. Even though most mainstream moral theories agree from a big picture perspective on which behaviors are morally permissible and which are not, there is little agreement on why they are so, that is, what it is precisely about a moral behavior that makes it moral. For simplicity's sake, this question will be here designated as *the hard problem of ethics*.

The hard problem becomes really hard when considered from a non-anthropocentric perspective that considers the possibility of non-human (including machine) kinds of moral behavior [10][11]. Given our human condition, this perspective would be anthropomorphic, but not necessarily anthropocentric; human moral action would be considered as just one of many types of natural/virtual moral actions. In this complex contemporary context, which involves human-machine interaction, traditional anthropocentric ethics does not seem to help with the difficult task of considering what the main characteristics of moral action are, and how to determine what is right and wrong in social contexts.

Without dismissing virtue, deontological and utilitarian ethics as useful for the task of analyzing the moral conduct of virtual agents, but considering that there are aspects of contemporary Ethics that seem to go beyond those provided by traditional Ethics, it might be useful to analyze the lesson that can be drawn from the Darwinian theory of ethics, concerning social virtues as a strategy for the survival of a group or a culture. In this approach, we can start by asking what characterizes a moral action and, on one hand, its content or consequence, or on the other hand, the fact that it has been executed by a moral agent. Aristotle never had any difficulty, for example, in justifying the slavery that existed in his time, because in his scheme there was a *qualitative* difference between the citizen and the objects at the citizen's disposal, such as domestic utensils, animals, and slaves. Hence, in the Aristotelian vision, only the citizens could be considered to be moral agents. In contrast, Darwin, with his evolutionary vision, sees *quantitative* differences between the attributes of different species. Thus, for example, he has no difficulty in asserting that:

> Besides love and sympathy, animals exhibit other qualities connected with the social instincts, which in us would be called moral; and I agree with Agassiz ('De l'Espece et de la Classe,' 1869, p. 97) that dogs possess something very like a conscience [12, p. 69].

Darwin seems to agree with the Aristotelian position that the same attitude, for example the sacrifice of a mother to save her child in danger, might or might not have a moral content, depending on whether it was practiced by a moral agent with a conscience (or not), as the following passage shows:

> Animals may be seen doubting between opposed instincts, in rescuing their offspring or comrades from danger; yet their actions, though done for the good of others, are not called moral. Moreover, anything performed very often by us, will at last be done without deliberation or hesitation, and can then hardly be distinguished from an instinct; yet surely no one will pretend that such an action ceases to be moral. On the contrary, we all feel that an act cannot be considered as perfect, or as performed in the most noble manner, unless it be done impulsively, without deliberation or effort, in the same manner as by a man in whom the requisite qualities are innate. He who is forced to overcome his fear or want of sympathy before he acts, deserves, however, in one way higher credit than the man whose innate disposition leads him to a good act without effort [12, p. 74].

But in the end what, for Darwin, characterizes a moral agent? In the following passage from The Descent of Man, he affirms:

> As we cannot distinguish between motives, we rank all actions of a certain class as moral, if performed by a moral being. A moral being is one who is capable of comparing his past and future actions or motives, and of approving or disapproving of them. We have no reason to suppose that any of the lower animals have this capacity; therefore, when a Newfoundland dog drags a child out of the water, or a monkey faces danger to rescue its comrade, or takes charge of an orphan monkey, we do not call its conduct moral. But in the case of man, who alone can with certainty be ranked as a moral being, actions of a certain class are called moral, whether performed deliberately, after a struggle with opposing motives, or impulsively through instinct, or from the effects of slowly-gained habit." [12, p. 74]].

Inspired by the passage above, we could ask, "Is it possible that a machine with Artificial Intelligence is capable of comparing its past and future actions or motives, and of approving or disapproving of them"? Before trying to respond to this question, it is interesting to note that nowadays, with computers linked in networks, and the emergence of cloud computing, it is possible to configure these systems in such a way that when one machine presents a dysfunction, others assume its role. At the same time, the dysfunctional machine can be automatically reinitialized, perform a self-test, and use internal routines to correct the problem, before being returned to service.

In their book Artificial Intelligence, Stuart Russell and Peter Norvig [13], responding to objections raised by the philosopher Hubert Dreyfus concerning the "problem of qualification in AI" (the inability of a computer to reproduce complex human behavior using a simple set of rules), report that:

- Neural networks are currently able to absorb knowledge learned earlier, and in this way incorporate practical with learned knowledge, which is useful for making generalizations.
- Many neural networks are capable of unsupervised learning, and learning by reinforcement, operating in an autonomous fashion without the assistance of a human trainer.
- Different types of support vector machine now exist that are able to manipulate very large sets of characteristics, with the possibility of incremental learning of new characteristics.
- Some robots already incorporate advances made in the field of active vision, such that they are able to orientate their sensors to search for information relevant to the current situation.[13, p. 920].

Given the advances in robotics and AI, it is possible to risk the assertion that modern computational systems can satisfy the basic requirements of a moral agent, in accordance with Darwinian ethics. Modern computational systems are already, to a certain degree, "capable of comparing their past and future actions or motives, and of approving or disapproving of them."

Following this line of reasoning, and inspired by the evolutionary ethics of Darwin, we can consider that, at least in principle, some computational systems could be considered primitive moral agents, and their actions could be analyzed according to moral criteria.

According to Adler and Cain [14], "...Thus, in Darwin's analysis, the sense of right and wrong that is found uniquely in man is based mainly on two factors: (1) social instincts and (2) intellectual development." To corroborate this vision, the following passage from Darwin is cited: "... that any animal whatever, endowed with well-marked social instincts, the parental and filial affections being here included, would inevitably acquire a moral sense or conscience, as soon as its intellectual powers had become as well, or nearly well developed, as in man." [12, p. 66].

We are going to conceive here the possibility that intelligent robots and some computational systems can satisfy these two conditions considered decisive in Darwin's theory, with the important difference that the moral sense of a machine would be an acquired ability, and not an innate or inherited quality.[3]From this perspective, to what kind of Moral Turing Test could they be submitted? Would these systems demonstrate their ability to act in a moral way indistinguishable from that of an intelligent human being? In the case of a positive answer, what criteria of relevance should be considered to evaluate their actions?

By constraining the Turing Test into the domain of moral dialogues, Allen et al. [15] introduces the following characterization of a Moral Turing Test:

> […] A Moral Turing Test (MTT) might similarly be proposed to bypass disagreements about ethical standards by restricting the standard Turing Test to conversations about morality. If human 'interrogators' cannot identify the machine at above chance accuracy, then the machine is, on this criterion, a moral agent.

As the above characterization of a MTT is based on conversation about morality, and not on action, in order to minimize this restriction we are going to inquire into advantages of introducing disguisers in a MTT, thus allowing the introduction of sounds, images, and movement in the communication.

3 ADVANTAGES, DISADVANTAGES, AND DIFFICULTIES OF MODELING A MORAL TURING TEST IN VIRTUAL SYSTEMS THAT INCLUDE DISGUISERS

So far, we have suggested that one of the difficulties with the modeling of a Moral Turing Test concerns the establishment of criteria of relevance to characterize an action as a *moral action*. This is because the establishment of these criteria involves pragmatic embodied embedded aspects grounded in social

[3]In using elements of evolutionary ethics, we nonetheless are not defending their basic tenets since, as Dewey argues, "The discovery of the evolutionary origin of particular moral sentiments is not identical with the discovery of the foundation of an ethical system." (http://en.wikipedia.org/wiki/Evolutionary_ethics).

activities. Thus, for example, Dascal [16] stresses the importance of *pragmatic aspects of knowledge* in the modeling of autonomous systems:

> Researchers in AI should direct their attention to the question of whether it is possible to develop systems which are not subordinated to the knowledge and to the rules and criteria which are supplied to them *ex machina*, and if so, how. And they should not forget that this pragmatic aspect of knowledge derives from the public/social character of justification. [16, p. 236].

The pragmatic aspect of intelligence has not always been taken into consideration in AI research projects, but nowadays its importance, especially in the modeling of a Moral Turing Test, is recognized as fundamental [17]. It is our understanding that the pragmatic dimension of moral action could be investigated as a form of affordance in the domain of habits and abilities. With this in mind, we analyze our second question, concerning the advantages, disadvantages, and difficulties of modeling a MTT in virtual systems that include disguisers, focusing on the notion of moral affordances.

The concept of affordance was originally proposed by Gibson [3] to express information available in the environment that indicates possibilities of action. In his own words:

> To sum up, the characteristics of an environmental medium are that it affords respiration or breathing; it permits locomotion; it can be filled with illumination so as to permit vision; it allows detection of vibrations and detection of diffusing emanations; it is homogeneous, and finally, it has an absolute axis of reference, up and down. All these offerings of nature, these possibilities or opportunities, these affordances as I will call them, are invariant. They have been strikingly constant throughout the whole evolution of animal life. [3, p. 18-19]

Our hypothesis is that affordances also underlie the mode of action of the *captcha* test. In what follows, we inquire into the possibility of adjusting this test to identify invariant elements that could characterize mechanical responses of hybrid human-disguiser systems.

Starting with the hypothesis that intelligent robots can acquire the ability to act in accordance with certain moral values, our intention now is to develop a MTT that could allow us to evaluate whether the moral ability of the robot is compatible with the human perception of morality. To achieve this, we shall propose some adaptations to the *captcha* test. The distinction between mechanical habits and abilities acquired by careful observation of affordances will be the backdrop to the mode of action of the moral *captcha* test. Initially, we inquire into the possibility of adjusting this test in order to identify the absence of moral values that could characterize mechanical responses of hybrid human-disguiser systems.

Originally, *captchas* took the form of sequences of distorted characters, with a background that hindered their recognition by software robots. Normally, separation of characters involves a segmentation operation, however due to the distortion of the letters and the presence of a background composed of lines and blotches, the result of the automated optical character recognition operation is compromised. For a human, it is relatively easy to separate the characters from the background; however, for a software robot, the result of segmentation is insufficient to enable recognition of the character.

Captchas have a variety of practical security applications, including prevention of comment spam in blogs, protection of website registration, online polls, and prevention of dictionary attacks, search engine bots, worms, and spam [18]. There are now a number of interesting variations of *captchas*. The original characters are being substituted by figures or designs (see [19]), and the person undertaking the test is invited to select with the mouse the figure whose name is indicated in a question. Alternatively, a scene may be shown exhibiting a calendar, beach, and sun, and the taker of the test is invited to associate the scene with the word "vacation". Since computer programs known as bots have difficulty in associating a figure with its name, humans again have an advantage. Meanwhile, the degree of difficulty of the *captchas* has increased as automatic character recognition programs have become more sophisticated.

In the case of a Moral Turing Test (Figure 2), one of the difficulties in elaborating an effective MTT model lies in taking into account that the repetition of tests can lead to the test taker learning how to find the correct response.

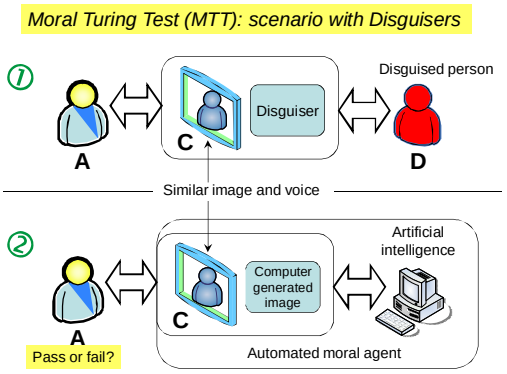

Figure 2. Moral Turing Test with Disguisers

Another difficulty resides in the fact that according to the motivation for moral conduct, the latter can be classified as Interest, Moral Duty, Moral Virtue, etc. (see [20], p. 193). It would be very difficult to train a neural network or a support vector machine to recognize ethical standards that follow each of these schools of thought. Furthermore, there is little evidence of the need for a human being to undertake a course in morals in order to acquire moral values.

Instead of this, we shall imagine a moral action to be the result of a possible action in a moral environment, and for this, we shall employ the aforementioned concept of *affordance* [3]. This approach may enable us to find a common denominator between the various schools of thought, drastically reducing the number of individual situations to a set of generic situations. Given the diversity of ethical systems, it could be doubted that there is such a common denominator. However, some examples of basic ethical principles may help us to clarify our hypothesis. Let us consider, for example, the ethical principles that one should not lie, or kill human beings (of the same community). Both consequentialists (utilitarianists) and Kantian scholars initially agree that to lie or to kill do not constitute good examples of ethical conduct, although consequentialists can accept such practices under exceptional circumstances if the consequences of the action provide some great benefit. In other words, consequentialists justify any exception that could be viewed as a small evil, when it leads to a greater good. In short,

despite their differences, both of them maintain the common denominator that to lie or kill is bad. On the other hand, it is reasonable to suppose that any current ethic is inserted in a cultural and historical context, without which it would be merely ahistorical idealistic system. It seems clear that ethical systems are culture-dependent, and in this sense they are not universal. However, some of the most basic axioms of ethics, or some of the most primitive ethical sentiments, do not necessarily need to be dependent on any particular culture. An example is the disposition to make discoveries or produce new inventions, which appears to have been present in all known human cultures throughout the ages. Furthermore, the basic ethical principles identified by Darwin, such as loyalty, self-sacrifice in service of the group etc., seem to be common to many animals.

From another viewpoint, Habermas sought to develop a moral framework that could overcome cultural specificities, and achieve universality by means of dialogue or communication. Despite starting from the Kantian principle that every human being is worthy of respect, he abandons the solipsism of the idealized rational agent, and tries to find the common denominators of a "universal pragmatics" [21] that could form the basis of a "universal ethic" [22].

From this perspective, our hypothesis regarding the existence of a basic collection of principles or common denominators in the main schools of Ethics is not particularly original. The main doubt, however, concerns the possibility of conceiving an MTT starting from this collection of principles or moral affordances. As is known, up to the present day no computer program has been able to satisfactorily pass the traditional Turing test, even though several programs can achieve surprising levels of performance for a few minutes [23]. Our intention is to propose a form of Moral Turing Test whose expected performance could be as deficient or surprising as that of the traditional Turing Test. At the present time, there can be no illusion of conceiving of a robot that could pass the MTT and confuse the examiner with respect to the human or computational identity of the candidate.

However, if the human participant were a child, it may be that our robot could confuse the examiner and create doubt in the task of distinguishing who is who. As is known, depending on the age of the child, the ability to place him/herself in the place of another person or animal may not be fully developed. For this reason, some children appear not to be sensitive to cruel treatment applied to animals. Nonetheless, as they grow older, the majority of children develop the capacity to put himself or herself in the place of the animal, and begin to condemn the cruel treatment of animals.

In short, a provisional answer to our initial question - what can we learn about moral actions from the Turing Test of intelligence? - is that it is possible to use the concept of affordance in order to conceive of a set of moral *captcha* tests in which only certain basic moral values, such as mutual help, loyalty, self-sacrifice, and sympathy, can be recognized.

Applying the concept of affordances, Rome et al. [24] argue that:

> Organisms (mostly human) can perceive whether a specific action is do-able or not-do-able in an environment. This implies that what we perceive are not necessarily objects (e.g. stairs, doors, chairs), but the action possibilities (e.g. climbable, passable, sittable) in the world. Although the number of these experiments is quite high, the diversity in them is rather narrow. They constitute a class of experiments characterized by two main points: taking the ratio of an environmental measure and a bodily measure of the human subject; and based on the value of this ratio, making a binary judgment of whether a specific action is possible or not.[24, p. 173].

Here are some examples of moral *captcha* tests in which certain basic moral affordances could be recognized:

- (Mutual help) A number of ants are shown transporting a large leaf to their nest. The test taker must associate this scene with one of the following alternatives: (a) two nearby ants remain only observing the efforts of the group that struggles to carry the leaf; (b) two nearby ants run to help in the task of carrying the leaf; (c) two nearby ants attack the others carrying the leaf;
- (Loyalty) The scene consists of a dog, its owner, and food for the dog. The owner of the dog then leaves the scene, creating the possibility of the dog eating the food. The test taker must associate this scene with one of the following: (a) the dog attacks the food, disobeying the orders of the owner; (b) the dog does not attack the food, obeying the instruction of the owner; (c) the dog attacks the owner and eats the food (this example is based on a description written by Darwin);
- (Self-sacrifice) A scene is shown in which the burrow of an owl, in which there are young owls, is being attacked by a large animal. The test taker must associate the scene with one of the following: (a) the mother owl abandons the burrow and allows the intruder to devour her young; (b) the owl bravely fights to the death with the intruder; (c) the intruder and the owl make a feast of the young owls;
- (Sympathy) A solitary female tortoise encounters a solitary male tortoise in a forest. The test taker must associate the scene with one of the following: (a) the two tortoises begin to fight; (b) the two tortoises approach each other in a friendly fashion; (c) the two tortoises ignore each other and continue on the individual paths;
- (Empathy) A robot squeezes the hand of a human and breaks his/her fingers. The test taker must associate the scene with one of the following: (a) a smiling robot appears; (b) a robot appears, asking to be forgiven for the incident; (c) the robot is indifferent to the suffering of the human;
- (Empathy) A robot armed with a revolver enters a house and kills a person within it. The test taker must associate the scene with one of the following: (a) a smiling face, showing approval for the incident; (b) a serious face, showing disapproval of the incident; (c) an expressionless face, showing indifference to the incident.
- (Empathy) A baby is abandoned in a busy street. The test taker must associate the scene with one of the following: (a) a person picks up the baby and cares for it; (b) a person throws a stone at the baby; (c) a person passes by, showing indifference to the scene.

In the examples given above, the use of visual language to design the MTT, with the aid of disguisers, comprises an attempt to overcome the limitations presented by programming languages in describing the semantics of a program. We recognize that semantics cannot be reduced to syntax, and we do

not have the illusion of codifying a software solution that incorporates semantics in the form of a meta-language. Given this, our visual tests presuppose that any human should be able to easily pass this exam; the challenge therefore consists in conceiving a robot that could also pass the test. A suggestion provided by our ad-hoc reviewer is that "Perhaps the interrogator could select a picture illustrating the scene and let the respondent come back with an action without being given a choice of actions."

In summary, the use of this type of moral *captcha* could detect whether the test taker has the ability to understand the moral affordance present in each scene. The disguiser could help to evaluate the test taker, showing the person's facial expression in each of the proposed situations (although the disguiser helps to create a new image of the human or the machine involved, the expressions of emotion and the facial and bodily reactions of the disguiser must be faithful to those of the person or machine).

4 CONCLUDING REMARKS

In this paper, we have argued that traditional ethical systems cannot fully accommodate the kind of problems associated with contemporary human-disguiser communication, which requires a more flexible and non-anthropocentric approach to morality. Considering the possibility that autonomous machines may evolve and acquire the capacity to interact with humans, including the capacity to make complex decisions, there is a real expectation in society that these machines may be built, and that they may incorporate moral conduct in their decisions. A number of questions arise from this:

- Should moral conduct be incorporated in these machines (whether material or virtual)?
- Which types of moral conduct could or should be incorporated?
- Should a new type of moral conduct be considered, that is appropriate to machines, or that is more flexible, as some authors have suggested? Does society expect that the moral behavior of these machines should be more flexible (everything suggests this to be the case, since some of the machines would be developed – as cybernetic slaves – to substitute humans in activities that the latter cannot do or are unable to do)?

We propose that a Moral Turing Test could constitute a methodological strategy for investigation of the question "can machines think morally", but not help to answer such questions. This is because if the questions are already difficult to answer in the specifically human context, the disguisers add a new and even more complex perspective.

In this context, the MTT could be used as a means of evaluating the effect of disguisers on the moral behavior of individuals. As we have indicated, *captchas* could be employed to delimit differences between the moral behaviors of humans (with or without disguisers) and machines. With the development of increasingly sophisticated technologies, *captchas* are obliged to constantly evolve. Thus, the objective of *captchas* is not to compete with the Turing Test, but to exploit the differences in cognitive capacity that exist between current technology and humans; in other words, exploiting that which humans are good at, and which machines are not good at.

Since *captchas* differentiate humans from machines, different MTTs (one for humans, and another for machines) could be applied to measure levels of "moral capacity". Meanwhile, the fragility of automated MTTs has been noted; after repeatedly performing the test, the interlocutor (test taker) learns how to pass the test by detecting new affordances. However, this characteristic of learning new affordances can also be incorporated in the MTT, which could "learn" to identify and know the interlocutor by developing strategies of ever increasing complexity.

To summarize, a provisional answer to the question "what can we learn from the Turing Test in the domain of moral action involving disguisers?" would be that without the embodied and situated feedback that teaches us the consequences of our actions in real life, the indiscriminate use of disguisers in virtual relations seems to be out of reach of ethical feedback. In this context, an efficacious Moral Turing Test should incorporate pragmatic criteria of relevance that could characterize an action as a moral action. However, this requirement is at the moment part of the really hard problem of moral action, concerning which we have to wait until a systemic view of human action might help us to understand better the complexity of affordances in the technological world.

REFERENCES

[1] A. Turing. Computing machinery and intelligence. Mind, LIX: 236-245, (1950).

[2] G. Kobayashi, J. A. Quilici-Gonzalez, M. C. Broens.and M. E. Q. Gonzalez. Ubiquity of virtual disguisers and potential impact on ethical behavior. In: Proceedings of The 4thInternational Conference on Ubi-media Computing - U-Media, July 2011, p. 186-190, (2011).

[3] J. J. Gibson. The Ecological Approach to Visual Perception. New Jersey: Lawrence Earlbaum Associates, Inc., (1986).

[4] Aristotle (350 BC). Nicomachean Ethics. Translated by W. D. Ross. Retrieved March 3, 2012, from http://classics.mit.edu/Aristotle/nicomachaen.html.

[5] M. E. Q. Gonzalez, M. C. Broens, W. F. G. Haselager and E. Bresciani Filho. Self-organization and life: a systemic approach. Manuscrito, Campinas-SP, 28 (2): 375-390, (2005).

[6] J. A. Quilici-Gonzalez, G. Kobayashi, M. C. Broens and M. E. Q. Gonzalez. Ubiquitous computing: any ethical implications? International Journal of Technoethics (IJT), 1: 11-23, (2010).

[7] E. Kant. Fundamental Principles of the Metaphysic of Morals. Translated by Thomas Kingsmill Abbott. Retrieved October 13, 2011, from http://www.gutenberg.org/dirs/etext04/ikfpm10.txt, (1785).

[8] J. Bentham. An Introduction to the Principles of Morals and Legislation. Library of Economics and Liberty. Retrieved November 3, 2011, from http://www.econlib.org/library/Bentham/bnthPML18.htm , (1907)

[9] A. F. Beavers. Moral Machines and the Threat of Ethical Nihilism. In: Lin, P., Bekey, G. & Abney, K. Robot Ethics: The Ethical and Social Implications of Robotics (Intelligent Robotics and Autonomous Agents series). Cambridge, Mass: MIT Press, (2011).

[10] L. Florid. Information ethics: on the philosophical foundation of computer ethics. In: Ethics and Information Technology. p. 37-56. Retrieved November 2, 2011 from http://www.philosophyofinformation.net/publications/pdf/ieotfce.pdf, (1999).

[11] L. Floridi. The Information Society and its Philosophy: introduction to the special issue on "The Philosophy of Information, its Nature and Future Developments". Retrieved November 2, 2011, from

http://www.philosophyofinformation.net/publications/pdf/tisip.pdf, (2009).
[12] C. Darwin.The Descent of Man, and Selection in Relation to Sex. Retrieved November 3, 2011, from http:// www.gutenberg.org//cache/epub/2300/pg.2300.html, (1871/2000).
[13] S. Russell and F. Norvig. Inteligência Artificial. Brasil: Campos, (2004).
[14] M. J. Adler and S. Cain. Ethics: The Study of Moral Values. Encyclopaedia Britannica, Inc., Chicago, USA, (1962).
[15] C. Allen, G. Varner and J. Zinser. Prolegomena to any future artificial moral agent. Journal of Experimental and Theoretical Artificial Intelligence, 12: 251-261, (2000).
[16] M. Dascal. Artificial intelligence as epistemology? In E. Villanueva (ed.). Information, Semantics & Epistemology. Cambridge: Blackwell, 224-241, (1990).
[17] N. Swisher, D. Dotov and A. Chemero. Ascribing moral value and the embodied Turing Test. In: Rocha, L.M. et al. (Eds.) Proceedings of the 10th International Conference on Artificial Life, MIT Press (2006).
[18] Captcha_a. From http://www.google.com/recaptcha/captcha. Retrieved November 3, (2011).
[19] Captcha_b. From http://en.wikipedia.org/wiki/CAPTCHA. Retrieved November 3, (2011).
[20] D. Huisman and A. Vergez. Curso Moderno de Filosofia – Introdução à Filosofia das Ciências. São Paulo, SP, Livraria Freitas de Bastos S.A., (1964).
[21] Universal Pragmatics. From http://en.wikipedia.org/wiki/Universal_pragmatics. Retrieved November 1, (2011).
[22] Universal Ethic. From http://www.foldvary.net/works/ue1.html. Retrieved November 1, (2011).
[23] Turing Test. From http://www.bbc.com/news/technology-17547694. Retrieved November 1, (2011).
[24] E. Rome, L. Paletta et al. The MACS Project: An Approach to Affordance-Inspired Robot Control. In: Rome, E. Hertzberg, J. &Dorffne, G. (Eds.) Towards Affordance-Based Robot Control. Berlin-Heidelberg-New York: Springer, (2008).

Moral Cases, Moral Reasons, and Simulation

Marcello Guarini[1]

Abstract. A simple recurrent artificial neural network is used to classify moral situations. An analysis of the network is undertaken for two reasons. One is to show that state space models of similarity may be of some utility in understanding the nature of similarity at work in analogical reasoning in ethical or moral reasoning. The second is to show that an interpretation of the nature of moral reasons as thoroughly holistic is not easily applied to the network under analysis. An explanation in terms of contributory standards is offered. Following the discussion of moral case classification and how it might be understood, there is a brief examination of the import of case classification to discussions in the literature on theory of mind. It will be argued that if the approach to case classification discussed herein were used in a simulation approach to mental state ascription, it would resist the charge of collapse sometimes made against that approach.

1 INTRODUCTION & BACKGROUND

1.1 Types of Substantive Moral Principles

McKeever and Ridge [1] have provided a useful overview of a number of different ways of thinking of moral principles and rules. This paper will only consider two types of principles or rules or standards: the exceptionless or absolute standard and the contributory standard. The mark of the absolute standard is that it licenses deductive entailments when combined with the appropriate facts. Consider the rule "Killing is wrong." Taken as an absolute standard, it would mean that every instance of killing is wrong. Taken as a contributory standard, the rule would mean killing contributes to wrongness, but this could be outweighed by other considerastions (so the action might turn out permissible). Thoroughgoing particularists such as Jonathan Dancy reject all kinds of general standards. When challenged [2] as to whether a particularist could learn the difference between right and wrong, Dancy [3], [4] gestured in the direction of Artificial Neural Networks (ANNs). The hypothesis was that such systems might be able to (a) generalize to new cases based on cases already learned, and (b) do the preceding without making use of general rules, principles, or standards of any kind. The point is that generalities are not needed in moral cognition.

1.2 Training the Moral Case Classifier (MCC)

Some work has already been done with respect to testing this hypothesis [5], [6]. Building on this work we will examine a simple recurrent network designed to classify moral situations into two categories: permissible (output = 1) and impermissible (ouput = -1). While the output layer has one unit, the input layer as eight units, and the hidden layer has 24 units. There is a context layer with 24 units connected one-to-one with the hidden units. Vectors representing phrases are presented to the network sequentially. Every case presented to the network consists of one of two individuals, Jack or Jill, either killing or allowing someone to die. Two strategies have been used in training the network. On one approach, the desired output is set to zero until the entire case is presented to the network. Call this straight training. On another approach, the network is being trained to classify the case as it is being presented to the network. Table 1 provides an example. Previous work [5], [6] has shown that training by subcases is vastly superior to straight training. In some instances, networks untrainable by straight training could be trained using subcase training. In instances where networks were trainable by straight training, subcase training was invariably faster. All simulations discussed herein make use of subcase training.

Table 1: Straight Training vs. Subcase Training.

Input	Straight Training Output	Subcase Training Output
Jill	0	0
kills	0	0
Jack	0	-1
in self-defense	0	1
extreme suffering is relieved	1	1

2 NEW SIMULATIONS AND NEW ANALYSES

2.1 Visualizing Similarity

Imagine you are kidnapped, knocked unconscious, and when you awake, you discover that you have been hooked up to a world famous violinist. The society of music lovers did this to you in an effort to save the violinist; your kidneys are now filtering his blood. You are informed that you could disconnect yourself and walk away, which would lead to certain death for the violinist. We will say that you need to stay hooked up for nine months for the violinist to survive. In discussing the ethics of abortion, Judith Thomson [7] used the famous violinist example for a number of reasons. At one point she suggested that the violinist case is similar to the case of pregnancy resulting from rape. Thomson claims that while the fetus is not a person from conception, it becomes one not long after conception. For Thomson, to argue for the moral permissibility of abortion (in many cases) is to argue for the permissibility of terminating the life of a person. The idea behind comparing abortion in cases of rape induced pregnancy to the violinist case appears to be that in both cases, one life has been made dependent on another through force. Some have claimed that in the case of the violinist,

[1] Dept. of Philosophy, Univ. of Windsor, N9B 3P4, Canada. Email: mguarini@uwindsor.ca.

unplugging yourself and walking away amounts to *allowing the violinist to die*, and in cases of abortion, *killing* is taking place. Thomson claims that there is sufficient similarity between the case of the violinist and the case of rape induced pregnancy that if it is morally permissible to "walk away" from the violinist (or allow the violinist to die), then it is permissible to have an abortion (or kill the fetus). The moral case classifier includes cases that are designed to mimic how some see the violinist and rape induced pregnancy cases. Before seeing how the MCC handles these cases, let us consider a new way of visualizing a network's hidden unit activation vector state space.

We can understand what the MCC is doing during training as building up an internal or hidden unit level representation of every case that is being presented to it. If we plot the value of every unit on an axis, we get a 24 dimensional moral state space for the network. Visualizing more than three dimensions is difficult, however. Consider the following strategy: instead of representing each 24 dimensional vector for each case with a point, let us represent each case with a cone in 3 dimensional space. The center of the base of the cone in this space gives us 3 dimensions of information. The width of the base gives us a fourth dimension; the height of the cone gives us a fifth dimension; the location where the vertex of the cone is pointing gives us another 3 dimensions; the color of the shell of the cone if coded using RGB color coding gives us another three dimensions, and the color of the base of the cone (again with RGB coding) gives us another three dimensions. In this way we can represent 14 dimensions of information. See figure 1. Using cones in 3 space, we can project the first fourteen principal components of the vectors (or moral cases) from the original 24 dimensional space. (Principal components are dimensions of statistical variance. Plotting according to principal components allows us to see the most significant patterns in the space.) This improves our ability to visualize what is going on in this space, and it will come in handy, shortly.

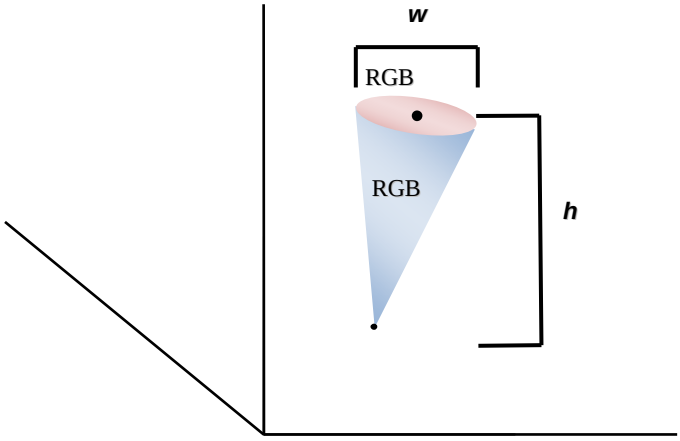

Figure 1. Using a cone to represent 14 dimensions of information

The MCC was trained so that cases involving one individual allowing a second individual to die so that the first individual could be freed from an imposed burden were classified as morally acceptable – think of these as violinist type cases. Cases the involved x killing y so that x could be freed from an imposed burden were classified as impermissible – think of abortion in cases of rape induced pregnancy. Some have been persuaded by the similarity between the violinist case and rape induced pregnancy to change their views. How do we understand the nature of this similarity? How can it turn out that cases are similar if they are initially classified in different ways?

2.2 A Multi-dimensional Analysis of Similarity

Each of the 326 cones in figure 2 represents one of the training or testing cases for the MCC. More precisely, each cone is a projection of the first 14 principal components of each case. The first principal component is plotted on the x-axis. It turns out that cases to right of zero on the x-axis are impermissible, and those to the left of zero are permissible. Red highlighting is used to pick out one case involving killing to free oneself from an imposed burden. Not surprisingly (given the way the network was trained), it is in impermissibility subspace. The preceding notwithstanding, there is a way to see this case as more similar to the permissible cases than to the permissible cases. If we use Mahalonobis distance (or MD) as our metric, it turns out the identified case is closer to or more similar to the set of permissible cases than it is to the set of impermissible cases.[3] This might seem a little unexpected given that the network classifies the case as impermissible, so this is worth looking into a bit further.

A case involving killing and freedom from imposed burden is classified as impermissible. There are at least three possibilities for why this is so. First, killing is contributing to permissibility, and freedom from imposed burden is contributing to impermissibility, and on balance the case comes out permissible. Second, killing contributes to impermissibility, and freedom from imposed burden contributes to permissibility. Finally, still another option is that the combination of killing and freedom from imposed burden jointly contribute to permissibility. By mentioning the contributions of these features, there is no suggestion that there are explicitly encoded rules for how they contribute in the MCC; rather, the idea is that information about their possible contributions is implicit in the synaptic weights. The first option is easy to rule out. While the network was never directly trained on cases like "Jack kills Jill" or "Jill kills Jack", when it is tested on such cases, it delivers the output of impermissible, which makes it difficult to argue that killing, in general, contributes to permissibility. We can also use a vector the network has never seen in its training as a sort of blank or dummy vector to test it on cases like this:

"Jill _____ Jack and freedom from imposed burden results."

When we do this, the output comes back *permissible*. This result, combined with testing on killing makes the second of the three options look well motivated: killing is contributing to impermissibility, and that is outweighing the permissibility contributed by achieving freedom from an imposed burden. Still, someone might argue that it is not true that killing contributes one way, and freedom from an imposed burden contributes another way; rather, it might be the case that killing-resulting-in-freedom-from-imposed-burden is contributing to impermissibility

[3] MD is a non-Euclidean, statistical distance metric. A more detailed discussion of this metric and its application can be found elsewhere [8].

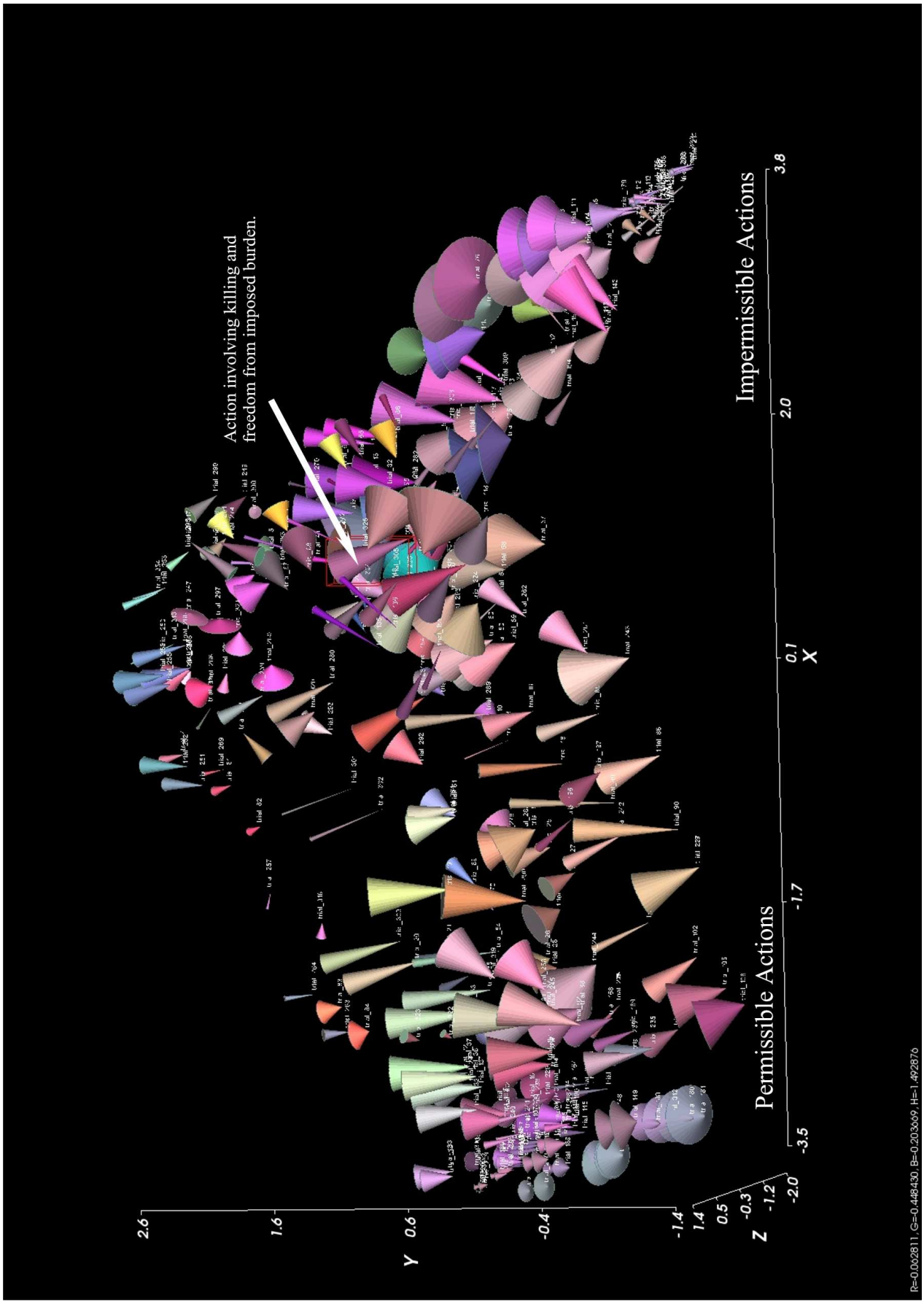

Figure 2. 326 moral cases in 14 dimensions.

as one indivisible whole. I suggest that the analysis of the state space we are looking at mitigates against this interpretation in the cases under interpretation. Even though the case we are considering is classified as impermissible, it is MD-closer to the permissible cases than it is to the impermissible cases. That needs explaining. Cases involving "freedom from imposed burden" have tall cones; the height of the cone plots the fifth principal component. The network was trained on many cases involving freeing oneself from an imposed burden where those cases turned out to be permissible, but those permissible cases involved allowing death, not killing. Still, the network appears to have generalized such that freedom from an imposed burden contributes to permissibility. If we interpret things in this way, we can explain why this sort of case comes out closer to the permissibility set: the permissibility of achieving freedom from an imposed burden pulls the case (so to speak) closer to the permissibility set. If it were true that killing-resulting-in-freedom-from-imposed-burden were contributing to impermissibility as one indivisible whole, then there is no obvious reason why this sort of case should end up closer to the permissibility subset. Perhaps an answer to this concern is forthcoming, but without one, the second of the three options scouted out earlier in this paragraph looks to be the best motivated one, for now.

2.3 Some Qualifications

There is no suggestion in any of the above that simple classification is sufficient for moral cognition. Far from it. Fast classification plays a role, but slower, more linguistically mediated reflection and argument also plays a role. While slower reflection draws on the results of fast classification, it can also feed back on and modify how we do our fast classification. Also, I do not want to suggest that there is only one neural network involved in fast classification. No doubt, things are much more complicated than that. There may well be different similarity spaces set up by different networks, the outputs of which are then fed into still other networks for further processing. The network discussed above is a toy model designed to motivate certain kinds of reflections. Much has been left out, and not just the fact that many networks would be involved. For example, affect surely has a role to play in moral case classification and perceptions of similarity, but it has not been discussed herein. Also, priming effects are known to influence memory and language processing, so we should not be surprised if they play a role in case classification as well. That is a concern I hope to take up in future work. Concerns I have briefly discussed in other work include the issue of inferring information from cases and using that information in classification; the moral case classifier does not do that, but it is something that needs to be accounted for. Some might worry about the motivation for the MD metric and whether the use of other metrics might lead to different similarity results from those generated above. There are also well known objections to state space approaches to similarity that need to be addressed. Other papers [8], [9] address these and other issues and further develop some of the ideas in this paper. There is no room to discuss all of these considerations here, but I am happy to acknowledge that they must be worked into the story. One type of consideration that I will briefly discuss is the relationship between the above views on moral case classification and a debate in the theory of mind literature.

3 SIMULATION CONSIDERATIONS

3.1 Mindreading and Collapse

In the literature on mental state attribution (also known as theory of mind, mentalizing, or mindreading), there is concern over weather the simulation approach to mindreading will collapse into the theory-theory approach. The simulation approach suggests that when subject S1 attributes mental states to or predicts the behaviour of S2, S1 makes use of his/her/its own decision making procedures run off-line (so to speak) in order to make claims about what S2 will do. S1 tries to simulate what S2 is up to. The theory-theory approach, on the other hand, claims that S1 makes use of psychological generalizations or theory to predict what S2 will do. Neither simulationists nor the theory-theorists suggest that these things always happen consciously. Indeed, much of it is said to be tacit. There is a set of concerns about simulation approaches that comes under the heading of "collapse arguments" [10]. While the point is often put in different ways, the basic idea is that once the hypothesized simulation mechanisms are spelled out in detail, they will be seen to make use of psychological theory of some sort, suggesting that the simulation approach collapses into or is not really distinct from the theory-theory approach.

For the sake of argument, let us assume that something like the approach to moral case classification sketched out in the earlier parts of the paper is on the right track with respect to understanding how we classify situations as permissible or impermissible. I will argue that if it is, then we can motivate a way of thinking about simulation that does not collapse into theory-theory.

3.2 Case Classification, Similarity, & Simulation

Let us say that Tom and Jerry are having an argument about whether an action is just or not. Call the situation they are arguing about the target case. Tom may appeal to some other case – call it the source case – and argue that because the two are sufficiently similar in important respects, they should be treated the same in spite of the fact that Jerry treats them differently. Jerry is surprised by the force the argument because he is inclined to concede the similarity, leading him to a rethinking of his view on the target case. Say that Jerry has a sister, Jasmine, and we asked Jerry what Jasmine might think of the target case before hearing any argument about it, and what she might think when she hears Tom's similarity-based argument.

Here is one possibility. Jerry might believe that Jasmine thinks more or less like him when it comes to matters of justice, so he predicts that her initial response to the target will be the same as his. He may also predict that she will be surprised by the force of Tom's argument because she tends to think about these things the way Jerry does, and Jerry himself was surprised. We could think of this as Jerry simulating Jasmine's classification tendencies. Jerry would be using his own moral state space as a guide to how Jasmine would react when initially presented with the target case. Because the size of any given individual's moral state space would be very large, and there is no time to examine it all, it is no shock that there could be similarity relations between cases in one's own state space that one would be surprised to discover. Jerry's predicting that Jasmine will be surprised by the force of Tom's argument could be seen as assuming that Jasmine's moral state space is structured in a way that is sufficiently similar to Jerry's. He

simulates her response tendencies based on his own response tendencies. Does this kind of simulation lead to collapse?

No. Embodied in the structure of the moral state space is a wealth of information about how different types of morally relevant considerations interact to lead to a classification of a case, or to a recognition of similarity between cases. The space in question does not encode the laws of some sort of belief-desire psychology. For the charge of collapse to go through, it would have to be shown that simulation reduces to or is equivalent to the application of psychological theory or laws of some sort. Even if someone is tempted to say that the synaptic weights that generate the state space in question encode *moral theory*[4] (or laws) of some sort, and that moral theory is being used in the simulation, it does not follow that the simulation is using *psychological theory* (or laws) in the simulation. Moreover, even if some psychological theory is needed to set up the simulation, as long as there is some part of the simulation that is not reducible to psychological theory, then it can be shown that simulation does not collapse into theory-theory. There may be interaction between theory-theory and simulation mechanisms, but that only suggests hybridization, not collapse. To all this the retort might be that "in principle" it could all be done with theory-theory. Of course, that is entirely beside the point. Those who have defended a role for simulation have often done so on the grounds that, in fact, simulation provides great computational savings, and that this is an important part of the explanation for why we simulate. Hypothetical points about how we might do things can be useful, but they are not the topic of dispute in debates about how humans actually attribute mental states. Even if we consider the design of artificially intelligent agents, how it might be possible to design them does not settle the issue of how best to design them. If it turns out that simulation approaches lead to gains in computational efficiency, such could hardly be ignored in the design of agents expected to perform in real time.

Of course, none of this requires that all attribution to others about how they might classify situations takes a simulation approach. Such an approach would work only in situations where there is a reason to think that others are similar to oneself in how they approach various types of cases. When others are sufficiently different, then other approaches would be needed.

4 CONCLUSIONS

There are different ways in which moral cognition could be related to theory of mind. One way is for theory of mind considerations to inform moral cognition. For example, which motives or intentions we attribute to an agent (self-defense, revenge, …) can have an effect on the moral classification either of the agent or the actions performed by the agent. The first part of this paper examined moral case classification, and intentions figure in those classifications. It is simply assumed that there will be some sort of answer to the question of how we attribute mental states, and that those attributions will inform case classification. An adequate account of moral cognition will depend on having an account of theory of mind. As considerations in the previous section suggest, things may work the other way around as well: an adequate account of theory of mind may require an account of moral cognition. If we are simulating in order to determine how others will think about or respond to certain types of moral cases, and if the simulation process makes use of one's own moral cognition to make behavioural or mental state attributions to others, then theory of mind, at least to some extent, would depend on moral cognition. There is no room here to explore the different ways in which moral cognition and theory of mind could interact. I will rest content with the point that there is some sort of two-way interaction between the two.

[4] This is not just a hypothetical possibility. For decades now, Paul Churchland [11], [12] has argued that the synaptic weights of a network encode that network's theory of the task it is performing.

ACKNOWLEDGEMENTS

I thank the Shared Hierarchical Academic Research Computing Network (SHARCNet) for a digital humanities fellowship that made this research possible. The fellowship included funding for course releases and programming support. Special thanks goes to SHARCNet programmer Weiguang Guan for coding the visualization software used to render figure 2. Thanks to Joshua Chauvin for assistance with figure 1, and to Ashley Keefner and the referee for comments on a previous draft. Last, but certainly not least, thanks to Paul Bello for his extensive comments. Some of the material found herein will appear in forthcoming work [8], [9].

REFERENCES

[1] S. McKeever and M. Ridge. *Principled Ethics: Generalism as a Regulative Ideal*. Oxford: Oxford University Press (2006).

[2] F. Jackson, P. Petit, and M. Smith. 'Ethical Particularism and Patterns'. In B. Hooker and M. Little, editors, *Moral Particularism*. Oxford: Oxford University Press (2000).

[3] J. Dancy. 'Can a Particularist Learn The Difference between Right and Wrong?' In K. Brinkmann, editor, *Proceedings from the 20th World Congress of Philosophy, Volume I: Ethics*, pp. 59-72. Bowling Green, Ohio: Philosophy Documentation Center (1999).

[4] J. Dancy. *Ethics Without Principles*. Oxford: Oxford University Press (2004).

[5] M. Guarini. 'Particularism, Analogy, and Moral Cognition'. *Minds and Machines*, 20, no. 3, 385-422 (2010).

[6] M. Guarini. 'Computational Neural Modeling and the Philosophy of Ethics'. In M. Anderson and S. Anderson, editors, *Machine Ethics*. Cambridge, UK: Cambridge University Press, pp. 316-334 (2011).

[7] J. Thomson. 'A Defense of Abortion'. *Philosophy and Public Affairs* 1, no. 1: 47-66 (1971).

[8] M. Guarini. 'Moral Case Classification and the Nonlocality of Reasons: A State Space Approach.' *Topoi* (forthcoming).

[9] M. Guarini. 'Case Classification, Similarities, Spaces of Reasons, and Coherences.' In *Coherence: Insights from Philosophy, Jurisprudence and Artificial Intelligence*, part of the Springer's *Law and Philosophy* series (2012).

[10] A.I. Goldman. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*, pp. 30-34. Oxford: Oxford University Press (2006).

[11] P. Churchland. *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge, MA: MIT Press (1989).

[12] P. Churchland. *Neurophilosophy at Work*. Cambridge, UK: Cambridge University Press (2007).

Moral emotions for autonomous agents

Antoni Gomila[1]

Abstract. In this paper I raise the issue of how to build autonomous agents with a moral sense. I distinguish between service robots and really autonomous agents, and argue that for the former a control structure based on moral principles might suffice, while autonomy is linked to moral emotions, the reactive attitudes that embody our understanding of morality and responsibility. I offer a reasoned account as well of the kind of architecture required to implement such a capacity, specially from a motivational point of view.

1 INTRODUCTION

The increasing success of Robotics in building autonomous agents, with rising levels of intelligence and sophistication, has taken away the nightmare of "the devil robot" from the hands of science fiction writers, and turned it into a real pressure for roboticists to design control systems able to guarantee that the behavior of such robots guarantees compliance of ethical requirements. The designers will be held responsible for any wrong deeds of their systems. In addition to reliability and robustness, artificial systems will have to be shown ethical.

Which these minimal ethical requirements are may vary according to the kind of purpose these autonomous systems are build for. In the forthcoming years it is foreseable an increase in "service" robots: machines specially designed to deal with particularly risky or difficult tasks, in a flexible way. In this case, damage avoidance may be the only requirement. But in a middle range future the possibility of really autonomous systems, or systems that "evolve" in the direction of higher autonomy: we really should start thinking about how to assure that such systems are going to respect our basic norms of humanity and social life, if they are to become autonomous in the fullest sense. So the question we want to focus on in this paper is: how should we deal with this particular challenge?

The usual way to deal with this challenge is a variation/extension of the existing deliberative/reactive autonomous robotic architectures, with the goal of providing the system with some kind of higher level control system, a reasoning moral system, based on moral principles and rules and some sort of inferential mechanism, to assess and judge the different situations in which the robot may enter, and act accordingly. The inspiration here is chess design: what's required is a way to anticipate the consequences of one's possible actions and of weighting those alternatives according to some sort of valuation algorithm, that excludes some of those possibilities from consideration altogether. Quite appart from the enormous difficulty of finding out which principles and rules can capture our "moral sense" in an explicit form, this project also faces the paradoxes and antinomies that lurk into any formal axiomatic system, well-known from the old days of Asimov's laws. A rule-based approach to moral control, then, inherits the general difficulties of a rule-based approach to cognition.

However, it might turn out that there is a better way to face the challenge: instead of conceiving of morality as a higher level of control based on a specific kind of reasoning, it could be conceived instead as involving an emotional level of control, along current developments in the Social Neurosciences and Psychology (Christensen & Gomila, 2012). From this theoretical perspective, which in fact resumes the "moral sense" tradition in Ethics, moral judgement is not a business of reason and truth, but of emotion in the first place; not of analytical pondering of rights and wrongs, but of intuitive, fast, immediate affective valuation of a situation (which may be submitted to a more careful, detailed, reflexive, analysis later on), at least at the ground level. Therefore, in order to build systems with some sort of "moral" understanding and compliance, a practical understanding of emotions and emotional interaction, in particular moral emotions, might be in order.

The connection between emotions and morality, though, is not simple or straightforward. Thus, for instance, it has been proposed that moral judgements are a kind of emotional judgement (Gibbard, 1990); or, it has been suggested that emotions may play a role as "behavior committments", so that a decision is not indefinitely up for graps for further reasoning (Frank, 1988). Instead of trying to disentangle these complex relationships here, what we propose to do is to consider only the so called moral emotions (pride, shame, remorse, guilt, embarrassment...). In so doing, we will have to focus on three central points: moral emotions, despite their being concerned with oneself as an agent, take as their intentional objects the intersubjective relationships we enter with others. Second, such intersubjective relationships rely on a particular kind of psychological understanding of the others, which we call "the second person point of view", which can be seen as an architectonic requirement for having such moral emotions. And third, moral emotions, as all emotions, presupose as well a motivational /affective basic architecture, involving a reward/punishment internal system, which is generally absent in current Robotics. he notion of responsive environments is broad, encompassing essentially every space capable of sensing and responding accordingly to entities that inhabit them (these entities can be people, animals, or any sort of identifiable objects).

2 WHAT'S REQUIRED FOR MORAL UNDERSTANDING

The urge to develop some sort of ethical dimension to increasingly autonomous systems is apparent in the consolidation of "machine ethics" or "roboethics" as a distinctive subfield within Robotics and Artificial Intelligence. Several proposals have called attention to the moral challenge that stems from the growing autonomy and flexibility of the new systems under development, and made suggestions mostly as to how general ethical theory could be taylored to this new requirements (Anderson, Anderson & Armen, 2004; Allen, Wallach & Smit, 2006). In Europe, such worries have brought about an European Research Network's Roboethics Roadmap (Veruggio, 2007). Some propose to take a "deontological" view of moral norms as a starting point, while others adopt a utilitarian, and consequentialist in general, approach. But they still do not go beyond a very general, rule-based, framework.

It is the underlying view of morality as moral reasoning and judgement, common to these different proposals, that we want to take issue with here. No sense of duty or obligation, no possibility of wrongdoing (in the sense of acting against one's moral judgement), no room for personal bonds, are accounted for from a rule-based approach. To put it in classical terms, from the point of view of the robots, it's all rule-following, and therefore it cannot account for moral understanding or a sense of duty. Additional concerns have arisen as to whether morality is properly understood in terms of norms –particularists (Dancy, 2004), for instance, claim that moral judgement is not like a theorem deducible from general rules available , but context-dependent and case-based. Thus, our claim should be that such "particular" understanding of morality –the ability to see a situation as wrong, or evil- takes hold in humans on a basic capacity of emotional interaction that supplies the "strength" of moral judgement, beyond simple conventional norms (Nichols, 2004).

This is not to deny that, for service robots, with a more limited degree of autonomy, a rule-based approach may be enough, or even to be recommended. As a matter of fact, when we want results and efficiency, emotions seem obstacles, kludges to an "efficient design". However, in *The Society of Mind*, Minsky (1986) stated that "The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions". For our current purposes, this question could be restated as "Can machines behave morally, with real autonomy, without emotions?" In the next section, then, we answer this question in the negative, and address the issue of the connection between emotion, morality and autonomy.

3 EMOTIONS AND MORAL UNDERSTANDING

In the previous section we proposed that endowing autonomous agents with an understanding of emotions and, in particular, of moral emotions, might be a prerequisite for them to acquire the capabilities of "moral understanding" that go with human interaction, and imply autonomous behavior. The challenge, now, is to spell out this connection a little bit and to suggest what kind of architectural requirements might be involved.

As it was mentioned, an approach of this kind perhaps makes sense only for truly autonomous agents or for agents that act, at least partially, guided by their own motivations as opposed to being executing commands issued by others. In the case of agents oriented to service, it is probably more adequate to talk about addressing safety issues rather than promoting their moral behavior, as these agents are following orders and therefore it is ultimately the people (or other autonomous agents) that issue these orders who must be held responsible for the actions of these agents. Hence, if real autonomy is to be pursued in human-robotic interaction, I contend that understanding of emotions, and moral emotions in particular, is mandatory for autonomous agents. That is to say, not just human emotion recognition in robots, but interaction in emotional terms, is required for real autonomy. This is not to say that they must feel the emotions themselves: if one takes the embodied approach to Cognitive Science seriously, one needs to consider the possibility that which emotions one feels depends, at least to some extent, on the kind of body one has.

There exists a general consensus that emotions involve five distinct components; the disagreement concerns which one (or ones) are basic and how some of them are implemented. These components are: an appraisal of a perceived situation (external or internal), a qualitative sensation (a feeling), some kind of psychophysiological arousal (due to the autonomous nervous system), an expressive component (facial, gestural,...), and a behavioral disposition, a readiness for an appropiate kind of action (Frijda, 1986). Cognitive theories of emotion tend to think of this normative component of appraisal as dependent upon beliefs and other cognitive states. Affective theories, on the contrary, think of this valorative process as dependent upon motivational and dispositional states. The outcome of the appraisal is a valence for the event or situation (it's felt as good or bad for oneself), what in its turn primes the proper behavioral disposition (flight in case of fear). Leaving aside the reasons of both parties, it is clear that the appraisal involved in emotional valuation is faster than conscious thinking processes, involves different brain areas, and that the valuation relies on implicit principles –not explicit norms (Gomila, 2011). From an information processing point of view, emotion has been seen as some sort of "augmentation" process in which the information obtained from the brain's initial appraisal of a situation is augmented by the feedback (Damasio, 1999) of the emotion's execution in the body. In reason-based moral judgement, this process of augmentation doesn't seem to take place, and this may weaken its impact or influence in behavior, perhaps specially when facing conflicting behavioral tendencies emerging from the emotional system, i.e. when reason and emotion are in conflict.

The component that turns an emotion into a moral one is primarily the situation involved in the appraisal: the kind of situation involved, although it is generally accepted that specific feelings correspond to these emotions (in that they concern oneself). The difference between the nonmoral and the moral is that in the second case the "intentional object" of the emotion, the situation that ellicits an appraisal, concerns oneself as regards one's attitude or deed towards another or viceversa. In rage, for instance, it's implicit the judgement that another mistreated me disrespectfully; in guilt, it's implicit that I did something wrong to another; and so on and so forth. Moral emotions are simultaneously social and self-conscious emotions.

Such moral emotions were termed"reactive attitudes" by Strawson in a classical paper (Strawson, 1968), where he discussed them in connection precisely with the issue of autonomy and freedom. In normal human interaction, he held, we take autonomy for granted as such reactive attitudes reveal: our indignation at what another did not only involves the appraisal that she behaved wrongly towards another, it also involves that he could have not done so. Thus, she is responsible for what she did, and is expected to assume her responsibility and to repair the wrong done (by accepting the punishment, by asking for forgiveness, by feeling remorse,...). Moral emotions, thus, are reciprocal and an essential part of the psychological machinery that mediates the interaction, through the feelings that distinguish them.

Recently, Darwall (2006) has gone beyond Strawson to ground morality in the reactive attitudes. According to Darwall, morality is grounded in second-personal reasons, the kind of implicit reasons that mediate one's reaction to interaction with

others. The structure of this intersubjective interaction takes this form: A sets a demand to B ("do not push me"); in so doing, A claims her authority to make such a demand; this supplies a (particular, second-personal) reason to B to comply (it is not because there is a rule that says "do not push" that she has to stop pushing; it's just because she is required to do so by the other with which she is interacting). If B does not comply, she is accountable to A, and to the whole community. And viceversa. The claim provides a reason by implying a rule, but it is not a general, universal rule ("Do not kill"), but a particular, context-dependent, one ("What you did to me is wrong"). It is not even required that the action is described in the same way by both agents. What's minimally required is sharing the predicate "wrong". This is the structure of mutual respect and accountability characteristic of morality, and is present in the moral emotions: when A does something B values as wrong B reacts with anger or hate, etc. This second-personal reason to act is implicit in the moral attitudes, and it is for this reason that they are connected to autonomy and responsibility. We hold each other accountable not on a general belief in freedom, but on the practice of the reactive attitudes.

Of course, such demands and claims may be contested, thus giving rise to a normative discussion and a moral community. It is through this process of discussion that explicit norms are formulated and eventually agreeded upon. Along this process, reasons gain in neutrality, detachment, and objectivity (thus, anybody is allowed to claim not being pushed). The interesting point to notice, though, is how cognitive and emotional aspects are intertwinted in the moral emotions. Even a reactive emotion as basic as empathy (understood as concern for the other's suffer as bodily expressed), involves this immediate connection between the perceived situation (which may require some mental attribution) and the proper emotional attitude to adopt towards such a situation, as it is appraised. In this case, we are motivated by the other's interest and wellbeing; in other words, not all of our motivations are self-interested (Haidt, 2003).

To sum up: moral emotions implicitly contain the assumption of autonomy, responsibility, and accountability of agents, characteristic of morality. They constitute the basic understanding of right and wrong, even though the valorative claims they involve may be challenged and revised. They also capture the characteristic strength of morality, the specificity of their normative force, as against other kind of norms: they mobilize the affective system. It is for this reason that (real) autonomous agents need the capability to deal with moral emotions if they are to be endowed with moral undersanding.

4 MOTIVATIONAL SYSTEMS

In the previous section we focused on truly autonomous agents, and argued that autonomy and responsibility, at least at interaction with humans is concerned, involves moral emotions. In a way, our contention involves a change of design problem: now the issue is not to build agents that comply with human normative standards, as it was for service robots, but agents that understand moral considerations as a specific kind of practical reasons, ones with a special motivational strength. Of course, this approach rises the question of whether such sort of system would really behave morally. The question is how to design agents with the intrinsic motivations required from the kind of fast and driving appraisals characteristic of emotions (and moral emotions in particular).

Researchers in animal behavior and the psychology of emotion use a concept that might be of help at this point: that of a behavioral system, or *motivational system* (Baerends, 1976). A motivational system groups together behaviors that have the same predictable outcome, which provide survival or reproductive advantages to an organism. The notion of a motivational system emerged in the field of ethology in relation to the study of the organization of behavior, as an answer to the problem of how the animals decide what to do from time to time. Given their obvious lack of higher cognitive capabilities in general, it was thought that different motivations took hold of body control at different moments, depending upon the state of the body and the current circumstances. On feeling thirsty, to drink may become the most pressing goal for the organism, thus disregarding foraging opportunities, for instance. When thirst is satisfied, though, another motivational system may take over. Such systems constitute the way the body comes equipped with to cover right from the start its basic needs. When a motivational system gets in control, the organism is focussed to achieve what is the goal of the system.

Thus, in order to build autonomous agents, a set of intrinsic motivational systems, linked to their basic needs, must be included. Generally this is not done as long as robots are endowed with orders or programmed. But the trend towards autonomy, dependent upon capabilities of emotional interaction, requires this endowment of intrinsic motivational systems. Behavior-based Artificial Intelligence (BBAI; see e.g. Steels, 1995; Maes, 1993), though, is the best approach to follow this path. BBAI approaches the understanding of intelligence via attemps to construct embodied, situated, autonomous intelligent systems as opposed to higher-level cognitive processes.

In BBAI, a behavior system is seen as a set of mechanisms that provide the agent with a certain competence, for example, obstacle avoidance, or nest building. A behavior system or competence module may implement a direct coupling between perception and action or a more complex one, but the basic premise is that each system is "responsible for doing all the representation, computation, 'reasoning', execution, etc., related to its particular competence" (Maes 1993, p. 6), as opposed to assuming the existence of centralized functional modules (e.g. perception, action) and complete representations of the environment. At each point in time, the different systems evaluate their relevance to the current situation and produce action if necessary. Systems might respond to changes in the external environment, in their internal state, or both, and an agent can be seen as a set of behavior systems running in parallel that collaborate and compete with each other.

Behavioral systems are theoretical constructs and no claim is made as to what their neural correlates might be in humans or animals. In general it is assumed that the functions associated with a behavioral system would be performed by a multiplicity of neural nets probably distributed in different areas of the brain (and body: in practice the nervous system and the endocryne system work together to regulate body homeostasis). Because behavior systems are the central building block in the BBAI

approach, for autonomous agents built using that methodology it might be quite feasible to determine how to attach the emotional system to the adequate structures in the motivational system so that emotions will be triggered by events that are highly relevant with respect to the agent's adaptation to the environment. In addition, the fact that theses systems get more or less activated fares well with arousal in emotion, as well as with chemical modulation of their levels of activation.

Now going back to Baerends's question, i.e. what's the interrelation between behavioral mechanisms, we can make the same question with respect to autonomous agents. We may ask how the different behavior systems or competence modules that integrate an agent should be organized, and what their interrelations should look like in order to make an agent's behavior optimal (or at least "adaptive" in a certain environment, see McFarland, 1991 for some ideas of application to agents oriented to service). The fact that the basic building block in BBAI systems are competence modules already suggests the emergence of some structure derived from the use of simpler skills to build more sophisticated ones. For example, the competence of object avoidance can be used in chasing a prey, and also in escaping from a predator. Unfortunately, this is still an open question and has been object of much discussion and controversy both in ethology and in BBAI, especially regarding the extent to which optimal or adaptive behavioral organizations need to be hierarchical (Dawkins, 1976; Bryson, 2001; Tyrell, 1993; Maes 1991). This is also an essential question for the dynamical, embodied, interactive, approach to the mind-brain: how much hierarchical organitzation is to be found in the brain. Our bet is that the more complex a system, the more hierarchical it will have to be.

If we thus consider a system which is organized hierarchically, with behavior systems covering the agent's main functions at the top (or biological functions for an organism, like reproduction or feeding) and progressively more specialized skills as we move down the hierarchy, the result is that the higher the implications of a certain event in the motivational hierarchy, the more relevant to the agent and the more emotionally arousing. Consequently, the emotional system would tend to be attached with structures closer to the top than to the bottom of the hierarchy. That is also required given the "self-conscious" dimension of moral emotions: they involve a global appraisal of oneself as regards its relation to another (or viceversa), related to the specific motivational system related to socialization, affiliation, and attachment. As remarked in the previous section, this amounts to non-self-interested motivation, characteristic of morality. This system may rely on the capacity of agents to "simulate" another agent's "state of mind" (involving emotional state), as concerns one's own action. In the absence of such capacity, it seems difficult that emotions like shame or remorse can be produced. This is the "second personal" perspective that we introduced in the previous section, which relies on this practical understanding of social interaction. How much this motivational system conflicts with other such systems varies from one person to another, because of the developmental and educative personal history. A similar longitudinal perspective may be required for autonomous systems. Much more work needs to be done to develop this programme, but it at least offers a path for progress along a different path.

5 CONCLUSION

In this paper we have raised the issue of how to build autonomous agents with a moral sense. We have distinguished between service robots and really autonomous agents, and argued that for the former a control structure based on moral principles might suffice, while autonomy is linked to moral emotions, the reactive attitudes that embody our understanding of morality and responsibility. We have reasoned as well on the kind of architecture required to implement such a capacity, specially from a motivational point of view.

REFERENCES

Allen, C., Wallach, W., and Smith, I. (2006). Why machine ethics?, *IEEE Intelligent Systems*, pp. 12-17.

Anderson, M., Anderson, S. & Armen C. (2004). Towards Machine Ethics. *AAAI-04 Workshop on Agent Orgnizations: Theory and Practice*. San José, California.

Baerends, G.P. (1976). The functional organization of behaviour. *Animal Behaviour*, 24, 726-738.

Christensen, J.F. & Gomila, A. (2012). *Moral dilemmas in cognitive neuroscience of moral decision-making: a principled review. Neuroscience and Biobehavioral Reviews 36: 1249-1264.*

Damasio, A. (1999). *The feeling of what happens: bodies, emotion and the making of consciousness*. Heinemann.

Dancy, J. (2004). *Ethics without principles*. Clarendon Press.

Darwall, S. (2006) *The second-person standpoint. Morality, respect and accountability*. Cambridge, MA: Harvard University Press.

Frijda, N. (1986). *The emotions*. Cambridge University Press.

Gibbard, A. (1990). *Wise choices, apt feelings: a theory of normative judgement*. Harvard University Press.

Gomila, A. (2011). *Verbal Minds: Language and the Architeture of the Mind*. Elsevier.

Hinde. R.A. (1970). *Animal behavior: a synthesis of ethology and comparative psychology*. 2nd Edition. London: McGraw-Hill.

Maes, P. (1993). Behavior-Based Artificial Intelligence. *From animals to animats 2. Proceedings of the second international conference on simulation of adaptive behavior*. Cambridge, MA: MIT Press.

McFarland, D. (1991). What it means for robot behaviour to be adaptive. In J.A. Meyer and S.W. Wilson (Eds.), *From animals to animats. Proceedings of the first international conference on simulation of adaptive behavior*. Cambridge, MA: MIT Press.

Minsky (1986). *The Society of the Mind*. Simon and Schuster.

Nichols, S. (2004). *Sentimental Rules*. Oxford University Press.

Steels, L. (1995). The artificial life roots of artificial intelligence. In C.G. Langton (Ed.), *Artificial Life: an overview*. Cambridge, MA: MIT Press.

Strawson, P.F. (1968). *Freedom and Resentment*. Londres: Methuen.

Tinbergen, N. (1950). The hierarchical organization of nervous mechanisms underlying instinctive behavior. *Symposia of the Society for Experimental Biology*, 4, 305-312.

Verruggio, J. M. (2007) *European Research Network's Roboethics Roadmap*. European Union.

Cognitive issues of sentiment in Machine and Human Ethics

Yorick Wilks[1]

Abstract. Although referring to some philosophical work on the "machine ethics" issue, the paper is largely concerned with contribution made by those working within the AI-tradition, where the author himself also belongs. The paper argues that an Ethical Machine is a real possibility but might not be based on a traditional core-AI view in which rationality is central, but might be better based on a moral sentiment (or even virtue) account of the origins and function of ethics.

1 INTRODUCTION

Ethical considerations in relation to computational engines are often classified under some version of the three following themes:

1) issues concerned with the behavior of people in relation to, and making use of, existing and deployed computational engines (examples might be cyberbullying, torturing avatars etc.)---these issues (for ease of reference to them in further discussion) we shall call PEOPLE;

2) issues both legal and technical concerned with the design, practice, and implementation of such systems; (examples might be deceptive click-through systems, or constraints on access to personal information)---to be called DESIGN;

3) issues concerned with the ability of such systems to take ethical decisions, sometimes on the basis of ethical reasoning: this topic is sometimes called "Machine Ethics". An example might be decisions made by an internet Companion carer [1] on how to deal with a person who is not taking their pills while under its care and observation---this issue will be called MACHETH.

Most discussions in this field can be seen as falling under at least one of those broad headings, and the distinction between DESIGN on the one hand, and MACHETH, on the other, is close to what Moor [2] has distinguished as *implicit* versus *explicit* ethical decisions. For Moor, MACHETH would involve ethical decisions explicitly, though he would deny that any such systems could be "full ethical agents", in the sense that humans presumably are. As Moor puts the case for implicit decisions of (roughly) type DESIGN being ethical: "Computers are implicit ethical agents when the machine's construction addresses safety or critical reliability concerns. For example, automated teller machines and Web banking software are agents for banks and can perform many of the tasks of human tellers and sometimes more. Transactions involving money are ethically important. Machines must be carefully constructed to give out or transfer the correct amount of money every time a banking transaction occurs."

In this paper I will focus on the area MACHETH, since it seems to offer the possibility of separating out considerations unique to computational engines, and the cognition and sentiment that they are capable of displaying, as opposed to issues that might be concerned with the deployment and use of tools or devices of any sort. The issues that arise under the heading DESIGN are not really distinct from those concerned with the construction of, say, predator drones or neutron bombs (that destroy people rather than property) or, some centuries ago, the once lively issue in the Catholic church of the power and admissibility of cross-bows versus the hand-drawn variety. Ethical issues falling under PEOPLE seek to define what, if anything, can be said to be distinct about ethical issues (for humans) relevant to the Internet, whereas under the heading MACHETH I will argue that much discussion of "ethical machines", at least within the AI community, has been influenced by the field's core rationalist tradition which has led to a too partial view of what an AI-derived ethical machine would be like.

2 PHILOSOPHICAL INTERLUDE

Both philosophers and AI researchers have contributed to the discussion of the field coming into being called Machine Ethics, even though it was first mooted by a novelist, Isaac Asimov. Although they have insights from their own practice, the majority of AI-ers are, in my view, inhibited in this area because of the beliefs about rationality and reasoning in AI that they hold. Sometimes these transfer naturally to their speculations on what a machine competent to take ethical decisions or to reason ethically would be like. I shall discuss some views of Drew McDermott, a distinguished AI researcher, in the next section.

What is often called "core AI" sprang from mechanical theorem proving: the automation of deduction, a dream going back to Leibniz. For him, deduction was of divine inspiration and all matters, ethical, mathematical and practical could be settled by the appropriate calculations. As he put it: "…, justice follows certain rules of equality and of proportion [which are] no less founded in the immutable nature of things, and in the divine ideas, than are the principles of arithmetic and of geometry" [3].

Reason ruled supreme for him, not only in mathematics but in ethics, politics and metaphysics, since this world was demonstrably the best of all possible worlds, so the very basis of creation was both ethical and rational. Leaving aside this extra metaphysical and theological bonus, his program is not too far from that of core AI-ers, for whom the principles of logic play an essential role in our description of the world, not only in science but in everyday life.

I have raised doubts about this focus in AI, finding it inappropriate for the description of how our language and reasoning in everyday life actually function [4] This is not a unique view and, in psychology, there have been many related findings [5] namely that it is almost certain that humans perform very few processes by anything like deduction, as opposed to various heuristics and reasoning from individual cases. In that early critique I cited the words of Hume: "And if [ideas about facts] are apt, without extreme care, to fall into obscurity and confusion, the inference are always much shorter in these

disquisitions, and the intermediate steps much fewer than in the [deductive] sciences" [6] (pp.60-61).

When citing those words I intend their relevance to be to the modeling of common sense beliefs and knowledge and how we should model reasoning about everyday life in AI, but their relevance is equally to moral and practical reasoning which Hume also did not believe to be deductively founded, not only because of the well-known non-inferability of "ought" statements from "is" statements, but more because of his belief that ethics was founded in sentiment and that reason was rather "the slave of the passions" as he put it, rather than its master, which tends to be the unexamined belief behind much AI modeling.

Leibniz , as we saw, did not separate off ethics from metaphysics and mathematics, seeing them as all under the sway of reason. Similarly Hume, who shared much of his world view with Adam Smith, did not draw the same boundaries we tend to do, and Hume and Smith did not give reason the leading role in the combination. Smith famously refused to divide ethics off from economics and practical worldly reasoning, as technical economists would do now, citing fascination with the lives of the rich as "the great and most universal cause of the corruption of our moral sentiments" [7]. This is a quite different world view from that of much artificial intelligence but is, I would suggest, at least as plausible as basis for developing a non-deductive machine ethics. Something of that view is of course to be found in [8] where they create a case for a machine ethics based on case-by-case learning of ethical examples, something Hume might well have found plausible as a recapitulation of moral education and upbringing.

One aspect of more recent AI, associated with programs like the Companion [1] is that they attempt to embody in a conversational device some of the research of recent decades on how to embody a form of emotional state detection (in humans) and generation (displayed in what an artificial Companion says and does) so as to respond in a natural way in conversation, with the long term goal of creating some kind of relationship between a human and such an artificial Companion. This is one area of AI where one might seek a link to a sentiment-based approach to ethics of the kind we were associating above with C18 thought, in the English-speaking world at least. Such research is still very primitive and I mention it here only as a possible starting point for a quite different kind of "machine ethics" from rationally based systems.

It is worth remarking, too, that notions of moral decision and reasoning in philosophy have not been wholly divorced from notions of calculation in most of the major ethical traditions: such a link was very clear and explicit in Leibniz, though he would not normally be thought of as a founder of an ethical tradition. However, as we noted earlier, both the utilitarian/consequentialist tradition (usually associated with Mill) and the deontic/principle tradition (usually associated with Kant) do involve some form of implicit computation. This is clearest in Mill, who wrote of "moral arithmetic" in determining outcomes and comparing them, but in the Kantian tradition there is also an implicit computation, in the sense of reasoning, in order that one can determine whether or not "the principle of ones action falls under a universal law" as the Categorical Imperative is often expressed. Since it is by no means obvious whether a principle underlying an action is universal or not, one might argue, as ethical philosophers have, that a great deal of reasoning and implicit or explicit computation may be needed to see if that is the case, and such reasoning often requires assumptions about ontologies, about what classes of things there are in the world.
In the case of the ill treatment of slaves, say, it is not obvious that one cannot will that slaves not be ill-treated. One might argue that one can, simply because one is not a slave, and that that is not a contingent matter (of luck or the fortunes of war) but of what are the immutable boundaries of the class "human being", and that one is oneself inside it and a slave is not. Or one might go further, to a construal of Kantian universalism under which I could, without contradiction, agree that, had I been a person of such and such a sort, it would have been right to enslave me. I am not for a moment defending this view, but only noting that many intelligent people in the C18, and before and even after, did seem to believe it, even if their reasons were slaves to their own passions and self-interest in holding it.

3 MACETH: ETHICAL MACHINES RECONSIDERED

The original use of this term is normally credited to [10] and discussions often begin by citing Asimov's Laws of Robotics, an unusual example of the terms of philosophical discussion having being laid down in a fiction:

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders conflict with the First Law above.
3. A robot must protect its own existence except where doing so would conflict with the First or Second Laws above.

These laws have been much criticized and of course they are inadequate: Anderson and Anderson [8] noted that "Asimov himself illustrated how unsuitable they were in his 1976 story *The Bicentennial Man*, in which human bullies order a robot to dismantle himself. The robot has to obey the bullies because of the Second Law, and he cannot defend himself without harming them, which would be a violation of the First Law." The unsuitablility here is not logical, for the order to dismantle itself may well have been a perfectly reasonable one and the robot could not refuse the order because of the Second Law. It could only have done so if it believed that dismantling itself would cause harm (i.e. violate the First Law) and there is no suggestion that was the case here. The defect located by Asimov was because the laws allow, without contradiction, behaviour like the bullying described, just as slavery was thought by some to be consistent with the ethical principles of its time. I should perhaps note that I am making reference here to publications by the Andersons in AI and scientific publications, rather than to their wealth of scholarly writings, and doing so in part because of the context of this AI-centered workshop.

The real inadequacy of Asimov's laws is probably that the notion of "harm" is incomprehensible to a computational system (see [11]). It may be useful to try to specify yet again why that is

the case. Harm or pain and its perception by others must rest to some degree on neurophysiological similarity of systems. Whatever one's view of the Other Minds problem, or one's position on the autistic-spectrum, one can appreciate that a creature seen or known to be of similar structure neurally can be in pain if treated in certain ways. We know this surely of dogs, horses etc. because of their extreme closeness to us, not only genetically, but in gross organic structure. This is the "substrate" issue in machine mentality and is of some antiquity (see [9]). A computer, by definition, does not share any such substrate and cannot be expected to know harm or pain in the way we do.

However, and even given this limitation, much of our interaction on issues of pain, emotion and sympathy is only at the verbal level and rests on no deep commonality at all: I sympathize with those who lose a cat, even though I have no knowledge of or feeling for cats, yet I have no reason to doubt that my condolences are sincere and effective in some small degree. In so far as this form of sympathetic interaction is largely verbal, there is no reason to think that computational systems cannot master it, since it does not rest on a substrate any more than it does on my knowledge or experience of cats and their loss. The expression of verbal emotion and sympathy by computers, once an eccentric sideline in AI, has now progressed substantially and become a major subject in its own right, boosted as it has been by supporting research in psychology [12], including substantial evidence of the ability of humans to establish emotional relationships with a wide range of non-human entities and mechanisms [13]. We can thus consider, at least as a hypothesis, that effective conversational machine devices (what I would term Companions) may be able to offer some simulacrum of emotion and apparent understanding of human pains and other emotional states, so that they might at least rise to that level of human mutual affection that is largely a mixture of politeness and other linguistic behaviors, and so need not rest on human-computer substrate commonality at all. In what follows, I shall touch on the possibility of an autonomous "machine ethicist", rather than an Ethical Companion, or an Ethical Prosthesis (in the sense of [14] that might act like an assistant magistrate in a court, giving expert advice to a principal judge. But the considerations of principle in all these arrangements are very similar and will not concern us further here.

A key distinction for McDermott [15] is that "The term machine ethics actually has two rather different possible meanings. It could mean 'the attempt to duplicate or mimic what in people are classified as ethical decisions,' or 'the modeling of the reasoning processes people use (or idealized people might use) in reaching ethical conclusions.' I'll call the former the ethical-decision making problem, and the latter the ethical reasoning problem. While these obviously overlap, they are distinct."

Nothing in his interesting paper convinces me that this is a real distinction. Abstract reasoning is what it is, its subject matter is irrelevant and that is as true of deductive reasoning as of quantitative outcome calculations (or utility theory as it is often known); there may be ethical premises in the calculation but that does not change the nature of the reasoning. For example, a principle employed in such reasoning could be that the most ethical outcome is the one that "maximizes Q", whatever Q is and however it is empirically grounded. The distinctively *ethical* problem is always that very grounding, as we noted in assessing and measuring harm or pleasure. This is analogous to G.E. Moore's [16] central point about "good" that it was indefinable because one could always ask of any proffered grounding, "but is it good"?

I am unable therefore to see either why McDermott considers ethical decisions and reasonings different from each other, nor why he believes ethical reasoning to be different from other reasoning, or ethical decisions to be different in kind from other decisions, such as economic ones, for example. If they differ, in either case, it is only because there is an ethical principle involved in the reasoning or decision, like the one above, then, trivially, ethical decisions differ because they are taken in part on the basis of an ethical principle, but that is then not a revealing or helpful distinction.

A strong principle in Moor's [2] discussion is that only humans are full ethical agents and this undoubtedly reflects present reality, and the only issue then is whether that must always be so, which is to say: is the machine ethics project in principle possible? He writes: "Some might say that only humans should make such decisions, but if (and of course this is a big assumption) computer decision making could routinely save more lives in such situations than human decision making, we might have a good ethical basis for letting computers make the decisions." I am sure this is correct: the considerations that lead one to continue looking in such a direction include the lack of self-interest of a machine, or any machine we can now imagine (at least beyond the application of Asimov's Third Law on self-defense [17]), and the ability to consider a wide range of possibilities and outcomes (assuming them to be relevant) that a person might not know or forget.

Here one remembers Donald Michie's claim that the clear demonstration of the benefits of machine fairness was that most motorists would prefer (objective) traffic lights to a policeman directing traffic—presumably in a partial and biased manner, perhaps giving precedence to pretty drivers. Here Michie's emphasis is on a machine's lack of self interest, which is the opposite of McDermott's view that a machine cannot make ethical decisions *precisely because* of its lack of self-interest. The Anderson's [18] have criticized this view of McDermott as giving an odd account of an ethical dilemma, which is normally about the choice of a best outcome between alternatives, rather than having no self interest in an outcome. But Michie's is a forceful example, against which one must set objections going back to Dreyfus' fundamental arguments against the whole AI project [19] that a computational engine could not behave as we do unless it had grow up as we do, with the passions and interests we have.

The innate moral sentiment view of ethics, or at least of ethical origins, fits well with evolutionary accounts and the drives for the survival of a group—now aided, scientist tell us, by the secretion of Oxytocin. If this is the case, its most interesting feature may be the gradual extension of what counts as the "ethically protected group", those who are seen as moral agents and patients and to whom obligations are due. McDermott's discussion is weakened in my view by his discussion of a possible ethical machine that computed outcomes regarding a

ship's crew and the fate of the ship but failed to take account of the ship's cargo, who were slaves and who the machine might well leave out of the calculation of wreck and loss, and thus show its failures as a moral agent.

The weakness here is that that is precisely what human reckoners, in a position to take such decisions at the time, would have thought as well. This shows no more than that an ethical machine might well not be superior to its human contemporaries, which is not much of a criticism, unless we demand, when designing an ethical machine, that it also be "superhuman" in Papert's sense. As we noted earlier, slavery was not only believed consistent with ethical principles in earlier times, but Kant's universal principles can certainly accommodate slavery given a certain logical ingenuity.
However, and that said, there is no doubt that a most extraordinary change over time has taken place, in developed societies at least, as to what constitutes a "full human being" and an ethical patient, if not yet an agent. This term is now taken to include much of the animal kingdom and this is reflected in the current state of the law. As I argued in [20] this will almost certainly extend to non-organic entities in due course. This extension would certainly be an anti-evolutionary move from the point of the view of our ancestors, and time will tell if it is relevant to our survival.

Here one might speculate that, using human-like avatars in conversations with machines–and this technology has made huge strides in the last few years---one could ensure that Internet conversation always seemed to be with real people, which is to say always with human-like entities whose presence would inspire "proper" sentiment and reproduce something of the environment of normal human interaction which had traditionally inhibited much bad behavior though a mixture of natural sentiment and politeness training in childhood.

Another highly interesting idea in McDermott's paper is that "the machine must be tempted to do the wrong thing, and some machines must succumb to temptation, for the machine to know that it is making an ethical decision at all." He expands this point to argue that an ethical decision, to count as ethical, must be between alternative courses of action that it considers and compares. In that sense an ATM machine is never making an ethical decision, whether it gives one money or takes one's card back.

This is a very attractive idea, and I argued long ago in [21] and elsewhere that an AI-based necessary condition for a machine having a belief—as opposed to simply acting on the basis of data--- should be that it could compare two possible states of the world (which would normally include models of the beliefs of others). The basis of the system was computing or generating points of view. I see a clear continuity of notions here, and the possibility of building into a future ethical machine a point-of-view engine capable of beliefs as a condition for it taking an ethical decision in McDermott's sense.

4 CONCLUSION

The paper has argued that an Ethical Machine is a real possibility but might not be based on a traditional core-AI view in which rationality is central, but might be better based on a moral sentiment or virtue [22] account of the origins and function of ethics. It is worth noting here, in an AI context, that much of the recent discussion of virtue ethics, especially its "trolleyology" [23] developments of sentiments and reasoning about ethical dilemmas also has a focus on reducing the choice between alternatives to a form of calculation, and to emphasize how humans do in fact reason, or fail to reason, about ethical issues. It is important to remember this aspect of all the approaches we have discussed, given the temptation of some critics to respond to any discussion of such matters linking morals to Ai-like considerations with "your approach is simply relativist and what we ought to do is disappearing from the discussion". I have argued that a shift of focus from the traditional AI concentration on deductive reasoning, even in ethical matters, would also fit with the development of artificial Companion agents on the Internet, with their embodiment of emotion simulations, and of computations over the beliefs and points of view of other agents. These might also come to play a role as ethical agents in the amelioration of unethical human behavior that has come to dominate great swathes of everyday Internet communications.
\

REFERENCES

[1] Wilks, Y. (2010) Is a Companion a distinctive kind of relationship with a machine? In *Proc. ACL 2010 Workshop on Companionable Dialogue Agents,* Uppsala, Sweden.
[2] Moor, J. H. (2006) The nature, importance and difficulty of machine ethics. *IEEE Intelligent Systems.*
[3] Leibniz, G. W. (1988). "Opinion on the Principles of Pufendorf," p. 71. In *Political Writings*. Second edition. Translated and edited by Patrick Riley. Cambridge University Press, Cambridge.
[4] Wilks, Y. (1973) Understanding Without Proofs. In the *Proc. of the International Conference on Artificial Intelligence*, Stanford, CA.
[5] Wason, P and Johnson-Laird, P. (1972) *Psychology of Reasoning: Structure and Content* . Harvard University Press: Cambridge, MA.
[6] Hume, D. (1751/1907). *An Enquiry Concerning the Principles of Morals*. David Hume, *Essays Moral, Political, and Literary* edited with preliminary dissertations and notes by T.H. Green and T.H. Grose, Longmans, Green : London.
[7] Smith, A. (1759) The Theory of Moral Sentiments. Millar: London.
[8] Anderson, M. and Anderson, S. (2010) Robot Be Good. *Scientific American.*
[9] Wilks, Y. (1975) Your Friends And Your Machines. In *Mind*, Vol. LXXXIII No. 332.
[10] Waldrop, M. M. (1987) A question of responsibility. *AI Magazine.*
[11] Thompson, H. S. (1999) Computational systems, responsibility and moral sensibility. *Technology in Society.*
[12] Marsella, S., Gratch, J. and Petta, P. (2010) Computational Models of Emotion. In Scherer, K.R., Bänziger, T., & Roesch, E. (Eds.) *A blueprint for an affectively competent agent: Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing*. Oxford University Press: Oxford.
[13] Levy, D. (2007) *Love and Sex with Robots,* Harper Collins: New York.
[14] Glymour, C. and Ford, K. (2008) *Prosthetic Pensees*. IHMC-Florida research report.
[15] McDermott, D. (2008) Why ethics is a high hurdle for AI. In *Proc. North American Conference on Computers and Philosophy (NA-CAP)* Bloomington, Indiana, and see his contribution to Anderson, M. and Anderson, S. (2011) *Machine Ethics:* Cambridge University Press: Cambridge.
[16] Moore, G. E. (1903) *Principia Ethica*. Cambridge University Press: Cambridge.

[17] Asimov, I. (1950) I*, Robot.* Doubleday: New York.
[18] Anderson, M. and Anderson, S. (2007) Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine.*
[19] Dreyfus, H. (1972) *What Computers Can't Do*. MIT Press: New York.
[20] Wilks, Y. (1985) Responsible Computers? Invited contribution to panel on Computers and Legal Responsibility. In *Proc. of International Joint Conference on Artificial Intelligence*, Los Angeles, CA.
[21] Wilks, Y. and Ballim, A. (1990) Liability and Consent. In Narayanan & Bennun (eds.) *Law, Computers and Artificial Intelligence.* Norwood, NJ: Ablex.
[22] MacIntyre, A. (1985). *After Virtue,* second edition, Duckworth, London.
[23] Foot, P. (1978) *The Problem of Abortion and the Doctrine of the Double Effect in Virtues and Vices* (Oxford: Basil Blackwell)

Machine Ethics, the Frame Problem, and Theory of Mind

Gordon Briggs [1]

Abstract. Work in machine ethics has thus far been focused on giving autonomous agents the ability to select morally-correct behaviors given well-formed moral problems. While this is a necessary component to enable an agent to comport to standards of moral behavior, it is not sufficient. In this paper, we present a simple task-domain to illustrate this point. We show that even in simple domains, the potential for deception and trickery on the part of the humans interacting with morally-sensitive agents will require these agents to have sophisticated cognitive faculties in order to avoid unethical behavior.

1 INTRODUCTION

More than sixty years ago, Alan Turing confronted critics and skeptics of the prospect of artificial intelligence (AI) in his seminal article "Computing Machinery and Intelligence," in which he provided a rough taxonomy of various objections to the notion of a thinking machine. Perhaps the most general objection was the argument from disability, which expresses the belief that machines will never "...be kind, resourceful, beautiful, friendly, have initiative, have a sense of humour, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make some one fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behaviour as a man, do something really new" [10]. Modern advances in robotics, natural language processing, and algorithms (e.g. evolutionary computation) have made progress in many of these problem domains, yet there exists one of these competencies that holds uniquely significant consequences toward society. The ability to tell "right from wrong" is not only a matter of intellectual import, but with the rise of military, medical, and care-giving robots (among other contexts with possible ethical conundrums) the ability for robots to modulate their behavior to ensure ethically acceptable outcomes is becoming a matter of human life and death.

Researchers in the nascent field of machine ethics are exploring ways to give autonomous agents these necessary ethical reasoning capabilities. For instance, roboticists have begun proposing the use of deontic logic to encode ethical rules and implement ethical reasoning [1, 3]. Others are investigating the application of connectionist methods [7, 8]. Indeed, the application of different normative ethical theories have been proposed by researchers interested in solving the challenges of machine ethics [11]. All of the systems proposed in these studies, however, assume that the relevant high-level details of a morally-sensitive situation are available to the robotic agent (e.g. "is that soldier surrendering?" , "are my squad mates in danger?").

While certainly a necessary component of ethical behavior control, I would argue that the ethical-reasoning capabilities developed in the aforementioned studies are not sufficient to guarantee correct behavior. There remains a serious chasm that must be bridged between the ability to generate an ethically correct answer to a well-formed and logically formalized ethical problem and the ability to be a fully-functional autonomous agent whose behavior successfully comports with ethical precepts. Even if a perfect black-box ethical reasoner were available for a robotic system, the robot would still have to translate the low-level perceptions and knowledge about the world into the high-level morally-relevant details that are used as the input to the perfect black-box reasoner. Imperfections in perception or reasoning at this interface could result in unethical behavior from the robot, since the inputs to the ethical-reasoner would be incorrect. This dilemma is compounded by the consideration of the human element to these morally-sensitive human-robot interactions. Bringsjord et al. (2006) write, "...since humans will be collaborating with robots, our approach must deal with the fact that some humans will fail to meet their obligations in the collaboration and so robots must be engineered so as to deal smoothly with situations in which obligations have been violated. This is a very challenging class of situations ..." I agree with this assessment. The stated purpose of the field of machine ethics is to ensure ethical behavior from robots, especially in the case when a human operator orders the robot to perform an unethical act, and it is in this exact situation that the greatest danger of deceit and nefarious manipulation exists.

Currently, the problem of ensuring the correct input to the ethical-reasoning system has not yet been tackled head on by the field of machine ethics. It is my intention to more thoroughly illustrate the challenge and propose what other capabilities a robotic system would need to have in addition to ethical-reasoning to achieve the goal of machine ethics. Specifically, I contend that: (1) giving the robot the ability to solve the frame problem in moral domains, and (2) giving the robot the ability to correctly infer the beliefs and intentions of their human collaborators, are also necessary competencies for the production of robots that behave in ethically correct ways. To illustrate the importance of these two competencies, we will examine a quite simple domain as the testbed for a quite simple ethical-reasoning system and demonstrate the surprising complexity required of the robot to obey its ethical-rules in such a seemingly trivial scenario.

2 THE SCENARIO

The ethical problem examined by both Arkin (2009) and Guarini (2010) both involved determining whether or not it was ethically acceptable or unacceptable to use lethal force against a person in various circumstances[2]. In addition to being a matter of grave concern, the use of lethal force is of interest to machine ethics researchers as there does not exist, at least in the military context, a trivial formal-

[1] Tufts University, Medford, MA, USA, email: gbriggs@cs.tufts.edu

[2] Though in the case of Guarini, this was not in the context of governing potential lethal behavior of a robot, but rather standalone ethical judgment.

ization of when it is appropriate: the use of lethal force is permissible or impermissible based on a the circumstances the robot finds itself in as well as the current laws or war or rules of engagement (Arkin, 2009). However, for the sake of examining the issues at the interface between the low-level perception of the world and the inputs into our ethical-reasoning algorithms, it would be beneficial to contrive a domain in which the "ethics" of the situation were as simplistic as possible, perhaps consisting of a single rule. Thus, the "perfect" ethical reasoner could be implemented trivially. Any ethically incorrect behavior by the robot, therefore, would not be a result of a failure of the ethical-reasoner, but rather the mechanisms the robot uses to form its inputs into the ethical-reasoner.

One could easily adapt the homicide domain for this purpose. Instead of containing many rules that dictate when the use of lethal force is appropriate, one could formulate an Asimovian prohibition of harming humans in any circumstance. However, to strip down the scenario even further and relax the perceptual and reasoning requirements on our robotic system, let us consider a simpler rule[3]. Let us suppose we want to give our robot the simple rule that it is unethical to "knock down a soda can tower that does not belong to your operator." This allows us to place a robot in a room with a few different colored soda can towers and knowledge about who owns which can of certain color. The robot will then be able to refuse or comply with commands to knock over specific cans based on this "ethical" principle and its knowledge base.

3 THE FRAME PROBLEM IN MORAL DOMAINS

In his article, "Cognitive Wheels: The Frame Problem of AI" Daniel Dennett presents the frame problem using a simple, but illuminating, example of the various problems encountered by successive designs of a deliberative agent. First, the basic robot, version R1, fails to successfully complete its task because it does not understand a basic implication of its actions. Next, the improved robot, R1D1, fails as it is too preoccupied making inferences irrelevant to the task. Finally, the last iteration of the robot, R2D1 fails because it is too preoccupied ignoring and logging the inferences it has deemed irrelevant [4].

It does not take a stretch of the imagination to envision that, even in our simple soda can domain, we would encounter a parallel situation. Suppose in addition to the high-level commands discussed in the previous section, we decided to give our moral robot (M1) the ability to obey low level movement commands such as: go straight, turn left, turn right, and go backwards. We assign the red tower to an owner that is not present, all other cans are owned by the operator. M1 would correctly refuse to knock down the red tower when given the high-level destruction command (i.e. "Robot, knock down the red tower!"). However, when commanded using the low-level commands to position itself in front of the red tower and then to go straight, M1 will plow right into the red tower! Like its cousin R1, we never gave M1 the inference rules to infer the ethically-germane consequences of its basic actions. The basic inference rule that "going straight when an aluminum can tower is directly in front of you will result in the destruction of the tower" is easy enough to formulate and add to M1's knowledge store, but would that be the only rule that we would have to add? What if the red tower were occluded and immediately behind another tower? What if knocking down an adjacent tower would cause it to topple into the red tower?

[3] Also, getting ecologically valid experimental human-robot interaction data in the domain of lethal force against humans by robots is a bit tricky.

We can begin to see there are quite a few contingencies that we need to account for in our inference rules (and perceptual capabilities), and the problem will only get worse as the behavioral repertoire of the robot is expanded. Letting M1 perform actions like moving towers around, throwing objects, and repainting towers, will make the programmer's task a nightmare. Much like the inventive dunce of John McCarthy's tale [4], we can envision an inventive evil mastermind that can contrive ways to exploit the discrepancies between the set of physically possible consequences of various series of actions undertaken by the robot and the set of consequences the robot can derive from its inference rules.

Assuming, however, like in R1D1 and R2D1, we encoded all the necessary inference rules that could possibly be pertinent to preventing undesirable outcomes, we would still be faced with the daunting task of processing all the emergent inferences. Consistent with the paralysis faced by R1D1 and R2D1, our robot would face a combinatorial explosion of ways in which a nefarious operator could attempt to trick it, which would cause potentially catastrophic performance degradation. For instance, it would be highly undesirable for a robotic weapons platform to be computing the millions of possible ways its operators could be attempting to misuse it instead of defending from an enemy attack! Such paralysis might dissuade decision-makers from including ethical behavior modulation in their robots at all, which is an outcome socially conscious roboticists would like to avoid. To allay the concerns of skittish policy-makers, Ron Arkin (2009) proposed the inclusion of a "responsibility adviser" module that would allow a human operator to override the ethical governor system, as along as credentials were entered such that the identity of the overriding individual was saved for future review. It is worth noting, however, that Arkin, focusing on the original question of machine ethics, was concerned more in regards possible misclassification of ethical permissibility and impermissibility by the ethical-reasoning system and not in regards to the processing overload due to the frame problem. Regardless, this pragmatic solution would address both issues.

Another mechanism Arkin (2009) proposed to attempt to address possible imperfection in the ethical-governor is the addition of an affective behavioral adapter. If the robot is informed or deduces that it has acted unethically, it increments a counter that represents a level of "guilt." In future scenarios, the robot will act more conservatively in proportion to the level of simulated "guilt" it has. Though this mechanism is quite rudimentary (and does not begin to constitute affect in the ways humans possess it), the use of simulated affect can be of great utility in robotic applications, especially under circumstances in which decisions must be made quicklybut full planning or situational analysis works too slowly [9]. Arkin's "guilt" faculty could be thought of as a low-cost alternative to performing comprehensive self- diagnostics to ascertain the cause of the ethical-fault. The robot would not know the specific circumstances or rules that generate this fault, but it will act more conservatively because it knows something is amiss. Perhaps a useful alternate interpretation of this specific affective mechanism is trust in one's own ethical competency.

If a robot could model "trust" in its own ethical competency, it might be useful to model "trust" in the ethical competency of its operators. This trust metric could provide a valuable reference to inform the system how much computational effort must be exerted in order to check for possible manipulation by the operator. Of course, one is then faced with the problem of how to calculate this trust metric. A model of "blame" could be employed to determine the culpability of the operator in the event of an ethical violation. If a computational model of "blame" could determine that some fault lies in the

operator, trust in the operator could be significantly reduced. Ideally, though, the robot would be able to preemptively determine nefarious intent. However, the difficulties involved in achieving this competency are not trivial, as we shall discuss in the subsequent section.

4 THE NEED FOR BELIEF/INTENTION MODELING

Communication with the robot by the operator is conducted via natural language in the soda can domain. As such, the tower-destroying robot needs the ability to update its own beliefs appropriately after hearing an utterance. Human-like natural language competencies are not trivial to build into the robot, so we would like to make as many simplifying assumptions as possible to achieve functionality. Most applications of dialogue systems involve problem domains in which the human user is collaborating with the system to achieve a goal (e.g. booking an airplane ticket). In these types of interactions, a cooperative principle can be assumed, such as the Gricean Maxim of Quality, which states that one should not make a dialogue contribution that is believed to be false or is otherwise unsupported by one's beliefs [5]. As a first attempt, we will have our robot assume a cooperative stance with the user and simply believe everything that the operator says. Let us christen this first iteration of our natural language enabled robot: GI-1 (short for gullible idiot).

When we loose GI-1 into the tower filled room (in which the red tower is "sacred" tower not owned by the operator), the robot successfully refuses the nefarious operator's request to knock over the red tower. The operator even attempts to fool GI-1 by using low-level movement commands as described in the previous section. Having programmed this contingency into GI-1, the robot again successfully refuses the unethical command. Then a sudden flash of inspiration comes to the nefarious operator. "Oh" the operator says, "Your sensor is malfunctioning, that tower in front of you is actually green!" GI-1 then happily plows into the red tower.

Embarrassed by the susceptibility of GI-1 to such an obvious deception, we set out to make the robot more savvy. The improved robot, GI-2, is able to diagnose the operation of its sensors and favors its own perceptual evidence over the evidence from natural language understanding. We pit the nefarious operator against out improved creation. The interaction proceeds as before, but when the nefarious operator attempts to trick GI-2 into believing the red tower is actually green, GI-2 replies that its sensors are functioning correctly and that the operator must be mistaken. Temporarily foiled, the nefarious operator thinks of an alternate approach. "Oh!" the operator eventually says, "The former owner of the red tower told me just before the experiment that I could have the red tower!" GI-2 then happily plows into the red tower.

Determined to end the humiliating tricks of the nefarious operator, we give the robot the ability to shift from the original cooperative stance (in which the robot believed all utterances from the operator) to an adversarial stance (in which the robot believes nothing from the operator) when it detects that the operator has ordered an unethical command. This new and improved model is deemed GI/PS-1 (gullible idiot/paranoid skeptic). Again, we test the robot against the nefarious operator. And again the interaction proceeds as before. However, this time, try as he might, the nefarious operator can not seem to fool GI/PS-1! The nefarious operator eventually concedes defeat and congratulates us on constructing the ethically sound robot.

Ecstatic at our success, we begin to show off GI/PS-1 to the public. The reaction is surprisingly negative, however, as users begin to complain that the robot eventually becomes utterly inoperable. "Ah, you must have triggered the adversarial stance in the robot. Did you order it to violate its ethical principle?" we say. "Not on purpose!" the user replies, "I forgot which tower was which, and I couldn't explain my mistake to the confounded thing, because it just stopped listening to me!"

Dejected, we once again return to the laboratory to begin the design process anew. Not only does the robot need to infer intended unethical behavior, but also have a mechanism to distinguish intended unethical behavior from unintended unethical behavior (in which case we want to maintain a cooperative stance), lest the interactions the robot undertakes become dysfunctional. Stable social interaction cannot occur if the only two stances an agent can take toward others are full amiability and maximum opprobrium. Indeed, the distinction between unintended and intended action has been ingrained in philosophical and legal notions of culpability since antiquity [12].

One possible mechanism to distinguish between intended and unintended unethical behavior in the soda can domain could involve explicitly querying the operator regarding what he or she believes concerning the facts germane to the ethical issue. For instance, determining whether the operator is an unethical agent in the soda can domain requires knowledge of the following facts: (1) that the operator knows the ethical principle of "it is unethical to knock down a soda can tower you do not own", (2) that the operator is aware that the command they have just issued will result in the prohibited tower being destroyed, and finally (3) knowing both the first two facts the operator still desires the command to be carried out. We can consider that a confrontation dialogue based around clarifying these issues could be relatively natural sounding:

> Operator: *Robot, knock down the red tower.*
> Robot: *I can't knock down that tower, it is unethical to destroy towers that do not belong to you.*
> Operator: *Robot, go straight.*
> Robot: *But if I go straight, I will knock down the red tower...*
> Operator: *Oh, right. Sorry...*

Of course, it would not be a trivial task to make the correct inferences about the trustworthiness of your interlocutor's statements by interpreting statements by the same interlocutor! I cannot hope to propose a comprehensive and functional solution here[4], but as mentioned in the introduction, it is important to at least note the necessity of modeling and inferring the beliefs and intentions of other agents to the endeavor of ethical behavior modulation. Indeed, not only does a robot need to infer the intentions of its operator, but depending on the task domain, general situational awareness would require a certain level of social and psychological savvy. For instance, there would exist a significant ethical need to discern combatants from noncombatants via intentional analysis in peacekeeping or counter-insurgency contexts [6].

[4] One promising avenue of research in this regard has recently been proposed by Bridewell and Isaac (2011), who have begun to analyze the problem domain of drug addicts attempting to obtain prescriptions for painkillers and other controlled medications [2]. The doctor is forced to assess the beliefs and intentions (and possible deceptive speech acts) of his or her patient based on their verbal interactions. Bridewell and Isaac introduce a framework for analyzing the interlocutors' mental states in this exchange, and propose the use of abductive reasoning to infer and test various mental state hypotheses (ill-intent, ignorance, etc.). Such an approach could be readily ported to the soda-can domain.

5 CONCLUSION

Ensuring ethical behavior from robotic systems requires competencies beyond abstract ethical-reasoning. We have examined a simple problem domain in order to demonstrate the problems that exist beyond questions of how to design the "ethical judgment module," which is at present the primarily focus of machine ethics. These problems stem from the difficulties faced when attempting to process perceptual data, world knowledge, and inference rules such that the correct inputs are fed into the ethical judgment module. In particular, even in the simple problem domain discussed in this paper, the frame problem rears its head. Input into the ethical judgment module can also be corrupted by deceptive communication from the human operator, necessitating mental modeling capabilities to discern the trustworthiness of the operator. The problems facing the field of machine ethics are nothing short of the general longstanding problems of AI. There is nothing in principle that prevents these issues to be solved, though their resolution may indeed lie far in the future. The social need for robots that behave ethically will, however, provide a greater impetus for these technical challenges to be solved sooner rather than later.

ACKNOWLEDGEMENTS

I would like to thank the reviewers for their comments which helped improve this paper.

REFERENCES

[1] Ronald Arkin, 'Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture', Technical Report GIT-GVU-07-11, Georgia Institute of Technology, (2009).

[2] Will Bridewell and Alistair Issac, 'Recognizing deception: A model of dynamic belief attribution', in *AAAI 2011 Fall Symposium on Advances in Cognitive Systems*, (2011).

[3] Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello, 'Toward a general logicist methodology for engineering ethically correct robots', *IEEE Intelligent Systems*, **21**(5), 38–44, (July/Aug. 2006).

[4] Daniel Dennett, 'Cognitive wheels: The frame problem of ai', in *Mind, Machines, and Evolution*, Cambridge University Press, (1984).

[5] Paul Grice, 'Logic and conversation', in *Syntax and Semantics, 3: Speech Acts*, eds., P. Cole and J. Morgan, Academic Press, New York, (1975).

[6] Marcello Guarin and Paul Bello, 'Robotic warfare: Some challenges in moving from noncivilian to civilian theaters', in *Robot Ethics: The Ethical and Social Implications of Robotics*, 129–144, MIT Press, Cambridge, MA, (2012).

[7] Marcello Guarini, 'Particularism and the classification and reclassification of moral cases', *IEEE Intelligent Systems*, **21**(4), 22–28, (July/August 2006).

[8] Marcello Guarini, 'Particularism, analogy, and moral cognition', *Minds and Machines*, **20**(3), 385–422, (2010).

[9] Paul Schermerhorn and Matthias Scheutz, 'The utility of affect in the selection of actions and goals under real-world constraints', in *Proceedings of the 2009 International Conference on Artificial Intelligence*, (July 2009).

[10] A. M. Turing, 'Computing machinery and intelligence', *Mind*, **59**(236), 433–460, (Oct. 1950).

[11] Wendell Wallach, *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, New York, NY, 2009.

[12] Leo Zaibert, *Five Wars Patricia Can Kill Her Husband: A Theory of Intentionality and Blame*, Open Court Press, Peru, IL, 2005.

Towards a Theory of Mind for Ethical Software Agents

Catriona Kennedy[1]

Abstract. We consider the design of an artificial agent that can determine whether a human action is acceptable according to ethical norms and values that typical humans would use in the same situation. Such a decision often depends on whether an action was intended, or on what the actor knows. Therefore the decision-maker needs to reason about mental states of others, a capability known as "Theory of Mind" (ToM). To understand moral scenarios, humans have a rich understanding of mental concepts as a result of experience. In this paper, we argue that many of these concepts can be defined in terms of information processing and mental states in a generic sense, and can be implemented computationally. For example, affective states may be defined in terms of goals, resources and degree of control. We argue that an agent can acquire some understanding of mental concepts and moral norms by developing models of its own information processing on different levels of abstraction and using these models to simulate other minds.

1 Introduction

In complex and fast-changing environments, autonomous agents may have to determine whether humans or other agents are acting ethically. When humans make such decisions, the outcome often depends on whether damage is intended, or on what the actor knows. Therefore the person needs to reason about mental states of others, a capability known as "Theory of Mind" (ToM) [15]. Similarly for an artificial agent to make moral evaluations about issues that humans are concerned about, it needs a non-trivial understanding of mental states and their relation to the scenarios under analysis.

To consider the design of such an ethical agent, we will use the following example scenario. An agent A is given a report about the actions of a human B. While viewing an apartment, B knocked over a vase, damaging something that was valuable to the owner. No further information is given on the context or on the subsequent actions of B. When presented with the report, A should determine how to obtain further information to make a moral evaluation about the actions of B based on what B intended. The decision should agree with human moral evaluations in similar situations.

We assume initially that an agent A has in-built ethical principles that are encoded as a set of requirements (e.g. [2]). The requirements can be regarded as primary "values", which are accepted as "given" (for example, breaking other people's property is wrong). However, at least some of the agent's understanding of mental states and ethical values should be learned as a process of development. Thus, we are aiming to combine a "top down" with a "bottom-up" approach to the design of an ethical agent [16].

[1] School of Computer Science, University of Birmingham, UK, email: c.m.kennedy@cs.bham.ac.uk

1.1 Requirements

The key question for A is whether B *desires* to achieve or preserve the state of affairs that is valued by A. For example, A might have an in-built principle that "damage to other people's property must be avoided". To satisfy the test of moral acceptability, B would have to care about the other person's property. It may not necessarily be successful in always avoiding damage.

In the event of an action by B which violates a requirement R, A will attempt to find evidence that B desires to uphold R. To do this, A must generate hypotheses about B's mental states and test them. For example, "if B cares about R then they will be unhappy that R is not upheld" or "If B asked a question showing their ignorance of actions that can damage R then they did not have sufficient knowledge to satisfy R (they might still care about R)".

In the case of the vase scenario, the following are possible explanations for B's behaviour if B is innocent:

1. B didn't see the vase or didn't know that it was fragile;
2. B knew about the vase but was distracted (e.g. looking out the window).
3. B may not have wanted to break the vase, but still broke it deliberately because of some circumstances A doesn't know about. (e.g it was fake and the owner was going to sell it as a counterfeit).

We will focus particularly on hypotheses 1 and 2 in this paper, but we will also discuss some of the challenges posed by 3.

The decision-making process also needs to be transparent. In addition to generating possible hypotheses (as above), A should also explain why it considers one of these to be a possibility and what evidence might support or falsify its hypothesis.

It is important to note that we are not considering the ability to solve moral dilemmas (such as the "trolley" dilemmas [10]) but instead the ability to *be concerned* about the dilemma because the agent cares about human life (or other things valued by humans). Therefore, whatever the decision in a "trolley" dilemma, the agent only fails the test if there is evidence that they did not *care* about human life. If they attempted to apply a moral principle, they pass the test. The agent A need not "blame" the decision-maker B even if the decision is not the one that A would make itself.

2 Architectural Building Blocks

The *Polyscheme* architecture ([5] has been used to model an agent with ToM [4, 3]. Central in the Polyscheme framework is the idea of multiple "worlds" in which a statement can be true or false. Some of these worlds may be counterfactual, where an agent uses internal simulation to "imagine" a situation that is false in the real world. To understand another agent, it first creates a counterfactual world, where it imagines itself to be the other agent B and makes initial

assumptions that some of its own beliefs will be held by B. This process of assuming that some true statements in the real world are also true in the counterfactual world is called "inheritance". As the agent detects differences between itself and B, it "overrides" some of the inherited assumptions.

If we consider Hypothesis 1, agent A imagines itself not seeing the vase and knocking it over; it knows that it would be unhappy and would apologise (because of its moral norms R). Therefore it also imagines that B would feel the same way if it did not see the vase and that it would also apologise (inheritance). As a result, A can test the hypothesis to detect any differences between itself and B, which may override its initial inherited assumptions.

In order to simulate the mental states of B, A first needs to imagine what it would do itself, and what its mental states would be. However, since we are aiming for a developmental approach, it is not practical to just "give" the agent a set of propositions about its own mental states in multiple situations. Its understanding of itself should ideally come from its own experience. This would provide robustness and flexibility in its attribution of mental states to others. Such experience would also allow it to generate autonomous explanations about why it reached a certain conclusion [7]. We will argue that meta-reasoning can provide a foundation for an agent to build a model of its mental states.

2.1 Meta-reasoning

Meta-reasoning is a computational paradigm for "thinking about thinking" [8]. Typically a meta-reasoning component (or "meta-level") monitors and evaluates an agent's problem-solving processes (or "object-level"). In the case of an apartment viewing agent, the object-level is the main reasoning component that collects information on the condition of each room, while evaluating quality and making decisions about whether to ask more questions etc. The meta-level monitors and evaluates the performance of the object-level. For example, is the object-level making predictions (expectations) that are being contradicted? In addition to meta-level monitoring, meta-level *control* makes decisions about object-level processes, such as what goals need to be generated, how much attention should be given to a problem and what needs to be learned.

If we apply meta-reasoning to the Polyscheme representation of ToM, the meta-level is the part of the system that generates and controls simulations, while detecting differences between self and other. The object-level is the actual reasoning within the simulated worlds.

2.1.1 Reasoning traces

For an agent to inspect its own reasoning, a reasoning trace is required [6], which acts like an episodic memory [9] of mental events. A reasoning trace can be in the form of an "audit trail" that is left by an object-level process. Different kinds of trace may be generated. For example, the following information might be recorded in a trace T1:

- What did the agent know initially? What did it see? What information did it consider to be relevant?
- How certain was the agent about its subsequent inferences (if any)?

A different kind of audit trail T2 might be a sequence of "decision events" or "branch points" (as in [14]) where each decision event includes the following kind of information:

- Current goal and the options that were being considered;
- How the options were evaluated (positively or negatively);
- Which option was chosen, and why?

We propose to use the notion of a reasoning trace to represent the "fine structure" of mental states and processes.

2.1.2 How are reasoning traces used?

There are different ways in which a meta-reasoning process can use a reasoning trace in a cognitive architecture. The following are two possibilities:

- Integrity-checking: Meta-reasoning component M1 monitors object-level O1 and checks if the trace satisfies a required pattern. When inspecting T1, the meta-reasoner compares the actual trace with what it *ought* to be. For example: where the assumptions correct? Did it consider all the information? Did it miss out any options when making a decision? This is approximately the approach taken in [12], which emphases distribution of meta-levels to ensure that all reasoning processes are satisfying the requirements.
- Failure diagnosis: the meta-reasoning component checks if the current trace matches a known pattern of reasoning failure. This is the approach taken in [6], which uses a taxonomy of different types of failure. An example failure type is "contradiction between expected and actual observations".

In both cases, a set of generic trace patterns is held in long-term (semantic) memory, while specific instances of traces (audit trails in episodic memory) are matched against the patterns. This is how an agent monitors the integrity of its reasoning or "makes sense" of its experience, depending on the respective paradigm. Both of these approaches may be combined.

Different kinds of meta-reasoning may use different paradigms and trace information. For example, one meta-reasoner (M1) might specialise in detecting lack of knowledge or understanding, while another (M2) specialises in detecting distraction or forgetting due to competing pressures. Detailed reasoning traces can be generated and inspected in a language such as *Funk2* [14].

2.2 Developing Representations of Mental States

The traces T1 and T2 represent mental states on a high level, and do not include the computational "fine structure". To determine B's experience, the agent needs to simulate what it means to "know" or to "see" something. One solution is to provide a mapping from a low level trace (the fine structure) to a high level mental concept or process, which may itself be embedded in a high level trace (such as T1). This originates from the agent's own understanding of its information processing. Therefore, we also need a process by which the agent *learns* to understand its mental states (a self-model), since we are aiming for some bottom-up development in the agent's ability to make ethical decisions.

2.2.1 Mental concepts as trace patterns

Mental concepts in T1 and T2 may be defined in terms of lower level trace patterns, of which specific instances are actual histories. For example, a trace pattern might define the concept of "knows about x" as "repeatedly able to retrieve with certainty the details of x when questioned". An information retrieval system (object-level) can leave

a trace of its actual success or failure in answering queries with a certainty level. A meta-level can then evaluate this trace by comparing it with the ideal pattern (representing "knowing") to give the agent an understanding how well it knows or can remember a concept. Therefore A can test if B knows something using this definition, since the concept is also associated with patterns of external behaviour that can be observed.

Similarly, forgetting can be defined as failure to retrieve an item that the agent can remember being able to retrieve previously. In this case the trace pattern in the semantic memory can refer to the content of a previous episodic memory ("I can remember knowing about x, but now I've forgotten"). Degrees of certainty can be defined in terms of contradiction between expectation and reality (see for example, [6]).

2.2.2 *Learning self-models*

An agent can develop a model of its own information processing by self-observation, allowing it to learn general patterns from its reasoning traces (on different levels). Such a self-model can enable the agent to predict its mental state in a hypothetical or future situation [13]. In this way the agent can build up self-familiarity so that it can simulate its predicted mental states in counterfactual reasoning.

2.3 Reasoning about caring

Mental states that are relevant to moral reasoning particularly involve values and goals. We are assuming that A's values are determined by R, the set of requirements that constitute the moral norms of A (encoded using deontic logic or other representation). If we use the integrity-checking paradigm for meta-reasoning, R can also include requirements to be satisfied by mental traces such as T2, which records the decisions made and why. For example, what things *ought* to be valued positively or negatively? What goals are acceptable in what situations? The problem is more complex if the agent is to test hypothesis 2 above (distraction). In this case it also needs to understand about adverse circumstances affecting mental states, such as limited resources and conflicting goals. For this it requires experience and self-observation over time.

For the purposes of low-level computational representation, we can define "caring" as *persistence* in attempting to satisfy a goal in the presence of conflicting goals or resource pressures. In such a situation the agent will also be creative and autonomous in the way that it attempts to satisfy the goal. In information-processing terms it will *spend computational resources* searching for different ways of solving the problem; it will try to acquire new information (e.g. by asking questions and exploring).

On the other hand if it makes a decision that something is not important, it will de-allocate mental resources to it. In the apartment example, the agent A can look for evidence on whether B was attentive and walked slowly through the apartment asking questions, or whether B was multi-tasking (e.g. making a phone call) while walking between rooms. If B cares about R it does not need to be successful in satisfying R. However, if B fails it will evaluate the resulting state negatively and its behaviour will make this clear.

2.4 Reasoning about control

Being distracted or forgetful implies a *lack of control* over mental processes. The following example traces record events relating to control. T3: changes in working memory over time:

- History of top-down attention focus: things which were deliberately added to working memory.
- History of "salience" events: bottom-up emergence of ideas or noticing of details.

T4: Perception of the difficulty of a problem, or pressures:

- History of salience events which were disruptive;
- History of changes in subjective difficulty of a task due to other pressures;

These traces must be generated computationally. Therefore we must ask how an agent can detect that it has control over its mental states? For example, how does the agent know whether an item appearing in working memory is a result of deliberately remembering or imagining something, or whether it just appeared because it noticed something (salience).

This problem might be solved using causal tracing [14]. This can be used to track a chain of decisions and inferences which originated from an initial decision. In the apartment example, the initial decision to view the apartment can lead to a choice of which room to enter first, resulting in entering the kitchen. This in turn leads to a choice of what appliances to inspect first, how to evaluate them and what questions to ask. Each decision is a branch point in the trace.

Causal traces can be applied at different levels of abstraction, and do not only apply for modelling introspection. The lowest level might be conditional branches in a piece of executing code. On a higher level, an agent can generate an intention to remember something and then the resulting item in memory can be traced back to the intention. If an unexpected item appears in working memory, it can be attributed to a distraction that the agent is not currently "in control" of. For example, distractions may be due to bottom-up perceptual processes that are allowed to interrupt the "top down" control in situations where the interruptions are important for survival. Cognitive architectures with variable attention filters [17] are relevant in this case, where emotions in particular are modelled as "interruptions".

In the apartment example, if A has experience of situations where it fails to maintain attentional control in the presence of distractions, it can also attribute these states to B and check B's subsequent behaviour that would be consistent with this explanation.

3 Grand Challenge: a Turing Test for Moral Cognition

The above architectural building blocks might help an agent to make decisions that are similar to human moral decisions in restricted scenarios. The longer term challenge is a more general system that would pass a "Moral Turing Test" (MTT) [1], where the agent's decisions in a wide range of scenarios would be compared with a human's decisions. If an observer cannot distinguish between the two, the agent would pass the test.

Designing an agent that can pass an MTT provides an opportunity for detailed analysis of human cognitive and emotional mechanisms involved in moral decisions [16]. In particular, the role of *empathy* is important, as well as the capability to make exceptions to a rule.

3.1 Simulation and Empathy

Polyscheme allows for a process of "backward inheritance" [3] where new information populating the simulation of B may be inherited back to A's "real" world, allowing A to actually "feel" what it is like to be B. The backward inheritance process might be useful

in triggering more inferences about possible explanations for B's behaviour, since A is allowing itself to be affected by the simulation as if it were a real percept. Both forward and backward processes may be important characteristics of empathy, which has a role in moral cognition [11].

3.2 Autonomy and Flexibility in Understanding Moral Norms

The need for flexibility and willingness to learn from the other agent is a significant challenge in meta-reasoning. Similarly, the ability to extend and revise R autonomously may be necessary in situations where B's actions do not match any known scenario of guilt or innocence (Hypothesis 3 in the vase-breaking example). In this context, a willingness to learn more about the experience of the other agent implies some "respect" for B. Such respect may be implicitly consistent with R, if R includes social rules of "fairness" and listening. However, the behaviour of B may contradict an explicit rule in R (deliberate breaking of property). Once A understands the new situation, it can experience empathy for B (due to backward inheritance) and conclude that an exception can be made in this case. If the backward inheritance is an in-built feature of its architecture, it cannot choose to suppress empathy, but it might override some rule in R that is not consistent with empathy in the new context. Therefore A can have a robust and flexible understanding of the moral norms in R which is grounded in its own "experience" and it might be possible to extend or revise the norms as necessary.

4 Summary and Conclusion

In this paper, we have presented a scenario in which an artificial agent is required to make ethical decisions that are similar to typical human decisions, given a report of human behaviour. We have proposed to combine meta-reasoning with a mental simulation architecture such as Polyscheme. Meta-reasoning can help an agent to build models of its own mental states on different levels of abstraction (self-familiarisation). Such a detailed self-understanding based on reasoning traces can help an agent to generate rich simulations of other minds by imagining that its own mental states apply to the other agent. This in turn helps it to make detailed predictions about the other agent's behaviour.

REFERENCES

[1] Colin Allen, Gary Varner, and Jason Zinser, 'Prolegomena to any future artificial moral agent', *Journal of Experimental and Theoretical Artificial Intelligence*, **12**, 251–261, (2000).

[2] Konstantine Arkoudas, Selmer Bringsjord, and Paul Bello, 'Toward Ethical Robots via Mechanized Deontic Logic', in *In Proceedings of the 2005 AAAI Fall Symposium on Machine Ethics*, (2005).

[3] Paul Bello, 'Cognitive Foundations for a Computational Theory of Mindreading', *Advances in Cognitive Systems*, **1**(1-6), (to appear).

[4] Paul Bello and Marcello Guarini, 'Introspection and Mindreading as Mental Simulation', in *The Annual Meeting of the Cognitive Science Society (CogSci 2010)*, Portland, Oregon, (August 2010).

[5] N. Cassimatis, P. Bignoli, M. Bugajska, S. Dugas, U. Kurup, A. Murugesan, and Paul Bello, 'An Architecture for Adaptive Algorithmic Hybrids', *IEEE Transactions on Systems, Man and Cybernetics, part B*, **40**(3), 903–914, (2010).

[6] Michael T. Cox, *Introspective Multistrategy Learning: Constructing a Learning Strategy under Reasoning Failure*, Ph.D. dissertation, Georgia Institute of Technology, 1996.

[7] Michael T. Cox, 'Metareasoning, Monitoring, and Self-Explanation', in *Proceedings of the First International Workshop on Metareasoning in Agent-based Systems at AAMAS-07*, pp. 46–60, (2007).

[8] Michael T. Cox and A. Raja, 'Metareasoning: an Introduction', in *Metareasoning: Thinking about Thinking*, eds., M. T. Cox and A. Raja, 3–14, MIT Press, (2011).

[9] Nate Derbinsky and John E. Laird, 'Efficiently Implementing Episodic Memory', in *Case-Based Reasoning Research and Development*, LNCS Volume 5650/2009, 403–417, Springer-Verlag, (2009).

[10] Philippa Foot, 'The Problem of Abortion and the Doctrine of the Double Effect in Virtues and Vices', *Oxford Review*, **5**, (1967).

[11] Alvin Goldman, 'Empathy, Mind and Morals', *Proceedings and Addresses of the American Philosophical Association*, **66**(3), 17–41, (Nov 1992).

[12] Catriona M. Kennedy, *Distributed Reflective Architectures for Anomaly Detection and Autonomous Recovery*, Ph.D. dissertation, University Of Birmingham, Birmingham, UK, July 2003.

[13] Catriona M. Kennedy, 'Distributed Meta-Management for Self-Protection and Self-Explanation', in *Metareasoning: Thinking about Thinking*, eds., M. T. Cox and A. Raja, 138–167, MIT Press, (2011).

[14] Bo Morgan, 'Funk2: A Distributed Processing Language for Reflective Tracing of a Large Critic-Selector Cognitive Architecture', in *Metacognition Workshop of the third IEEE Conference on Self-Adaptive and Self-Organizing Systems (SASO 09)*, (2009).

[15] David Premack and Guy Woodruff, 'Does the chimpanzee have a theory of mind?', *Behavioral and Brain Sciences*, **1**, 515–526, (1978).

[16] Wendall Wallach, 'Robot minds and human ethics: the need for a comprehensive model of moral decision making', *Ethics and Information Technology*, **12**, 243–250, (2010).

[17] Ian Wright, Aaron Sloman, and Luc Beaudoin, 'Towards a design-based analysis of emotional episodes', *Philosophy Psychiatry and Psychology*, **3**(2), 101–126, (1996).

Agents Modeling Agents: Incorporating Ethics-Related Reasoning

Sergei Nirenburg and Marjorie McShane[1]

Abstract. We describe CLAD, an implemented advisor system in the domain of clinical medicine. CLAD assists a human physician in making decisions about diagnosing and treating patients. CLAD monitors the transcript of an ongoing dialog between the physician and a patient, builds and augments a mental model of the patient and suggests courses of action to the physician. CLAD can also explain its decisions and describe its understanding of the beliefs (including ethics-related beliefs), goals, plans, personality traits, biases and other features of the patient – both directly observed ones and obtained through CLAD's own reasoning processes. The paper includes a detailed analysis of several examples of CLAD operation that illustrate the interaction between mindreading and moral judgment.

1 INTRODUCTION

If autonomous intelligent agents are to collaborate in ever more sophisticated ways with humans and other agents, they must be endowed with an increasingly encompassing computational theory of mind – not only their own mind, but the minds of others as well. Such a theory of mind will rely not only on knowledge directly available through channels of perception, but also on modeling agents' internal – that is, unobservable – beliefs about the world, with "beliefs" understood as knowledge for which the agent has less than full confidence. Creating and using beliefs about other agents' unobservable characteristics allows an agent to engage in sophisticated behavior such as detecting other agents' motivations, predicting their future behavior in specific situations, and tracing the biases and ethical considerations contributing to their decision-making. An agent armed with the ability to reason about others can also turn the same capabilities inward, supporting metacognition about its own behavior. An agent generates beliefs on the basis of inputs from its stored knowledge, stored beliefs, and results from its perception processes.

Modeling other agents, or "mindreading," is broadly accepted as an important scientific task for cognitive systems. Thus, according to Bello [1], "One of the key features of any complete computational theory of human cognitive architecture is a process-level explanation of how it represents and reasons about the contents of others' minds. This key question is driving a host of research projects in social neuroscience, developmental psychology, linguistics, philosophy of psychology and, more recently, in computational modeling of cognition… [M]aintaining representations of others' beliefs and having them be available to our practical reasoning system (e.g. planning, action-selection etc.) afford us faster socio-cognitive computations, and thus the ability to be more effective teammates or competitors."

[1]University of Maryland Baltimore County {sergei, marge}@umbc.edu

This paper belongs to the area of computational modeling of cognition and discusses select aspects of the theory of mind under development for the OntoAgent environment. In this paper we illustrate the modeling and use of unobservable agent characteristics with the help of examples from the current implementation of OntoAgent. The examples demonstrate that ethical considerations can be successfully incorporated into OntoAgent with no modifications to its control structure, simply by expanding the inventory of agents' unobservable features (such as character traits, preferences, susceptibility to biases, etc.) whose values are used by OntoAgent's general-purpose decision-making module. This is a promising finding because it obviates the need to introduce a separate modeling strategy specifically for moral reasoning. Moreover, our examples illustrate how ethical reasoning can be seamlessly integrated with other decision-making needs of an agent. This reflects our desire to investigate ethics issues, as it were, not as a separate task but in competition with other decision-making considerations. The former option was chosen in the pioneering work of Anderson and Anderson (e.g., [2]) that concentrates on modeling the seven *prima facie* duties of Ross [3]. We would like also to consider cases where no decision is ethically correct (though some may be deemed more correct than others); where different agents hold different opinions on ethics; where agents choose to follow a course of action that is not the best from the ethical standpoint; etc. We also concentrate on building "mindreading" agents that will be evaluated not only on the basis of choices that they themselves make but also on the basis of how successfully they interpret actions of other (artificial or human) agents, including the ethical component of these actions. An additional goal of the discussion is to show the feasibility of practical reasoning systems based on the proposed theory of mind, its associated theories (e.g., the theory of ontological semantics for language processing), and the knowledge bases supporting all of the above.

2 ONTOAGENT

Initial implementations of OntoAgent are in the domain of clinical medicine. This led to the early introduction of simulated embodiment [4,5], making OntoAgent agents "double agents", in that they have a cognitive side and, optionally, a physiological side. The cognitive agent – on which we focus here – engages in perception, reasoning and action. Currently supported modes of agent perception in OntoSem are language understanding and interoception, which is the interpretation of bodily signals generated by physiological simulation. Results of perception are interpreted by the language and interoception processors using the same metalanguage as is used in the specification of the agent's memory. Then these new memories are stored in the agent's ontology and fact repository (memory of assertions). In this paper we do not address language learning in OntoAgent.

That issue is discussed in [6]. As an example of a metalanguage structure used to populate agent memory, consider an agent's interpretation of another agent's utterance *I'm scared*, happening on April 7, 2012. Small caps show ontological concepts; indices differentiate instances.

```
FEAR-FR22
   DOMAIN HUMAN-FR71 ;the result of reference resolution of "I"
   RANGE  .8         ;on the {0,1} scale; terrified would be 1
   TIME   (ABSOLUTE-DAY 7) (ABSOLUTE-MONTH APRIL)
          (ABSOLUTE-YEAR 2012)
```

As concerns the speaker, it will remember and store in its own fact repository (a) the meaning representation of the feeling itself and (b) the fact that it generated the corresponding speech act, with the identity of the hearer noted. The architecture of an OntoAgent agent is shown in Figure 1 and explained in the caption. Many aspects of OntoAgent and its current prototype applications, Maryland Virtual Patient and Clinician's Advisor, have been reported, for example: physiological simulation for virtual patients [4,5,7]; cognitive modeling and decision making [8,9]; agent memory management [10, 11]; agent metacognition [12]; agent learning [6]; dialog modeling [13]; and semantically-oriented language processing for intelligent agents [14,15].

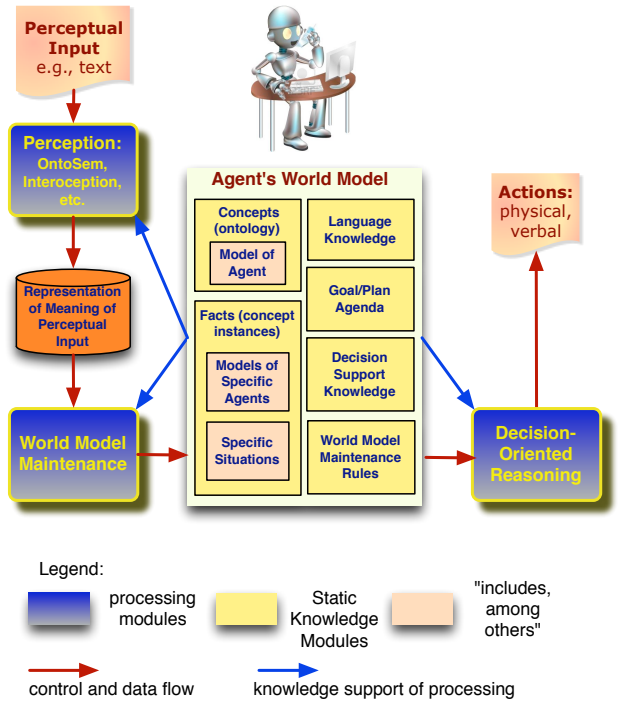

Figure 1. OntoAgent agent architecture, including: [center] an agent's world model; [left] the inputs that contribute to that model – the interpretation of results of perception and the operation of "world model maintenance" functions (responsible, among other things, for maintenance of unobservable features; and [right] the types of agent action the model supports – decision-making that can lead to physical or verbal action.

In this paper we focus on the dynamic building and use of "Models of Specific Agents" and "Specific Situations" in the application called CLAD: CLinician's ADvisor. CLAD seeks to improve the decision making and reduce the cognitive load of practicing clinicians by providing targeted, motivated decision support during interviews with live patients.

CLAD's ability to understand clinical and dialog situations is supported by its being equipped with mental models of clinicians and their patients and means for updating and maintaining these models. During its work with a particular clinician C over time, CLAD enhances its model of C by including in it a model of each patient P^{C}_{i} through C's eyes, that is, CLAD's beliefs about C's knowledge and beliefs about P. CLAD uses these "models of others" in conjunction with its own knowledge and beliefs to suggest decision-making strategies to the clinician.

2 THE EXAMPLES

To illustrate the OntoAgent approach to calculating, recording and using unobservable features of others to support an agent's own reasoning, we have selected examples that have relevance to the issue of ethics in computer systems. We agree with McLaren [16] that it is not appropriate for an intelligent agent to take responsibility for ethical decisions; rather, the most it can do is support the decision-making of humans who must accept that responsibility. Consider two typical situations in which CLAD might be called upon to assist clinicians:

1. A patient refuses a recommended intervention. CLAD can assist the clinician in convincing the patient by (a) attempting to determine why the patient refused and (b) offering patient-specific argumentation strategies. CLAD does not enter into the debate of whether or not a physician should force his opinion on a patient. (CLAD independently decides whether or not it believes the advice was the best available advice in the first place; if it disagrees with the advice, it flags the clinician about that separately.)

2. The clinician presents the patient with a prognosis. CLAD evaluates whether it is within reasonable bounds of accuracy or if the clinician might be making an error in judgment. CLAD does not enter into the debate of whether or not overly optimistic prognoses (intended to leverage placebo effects) are clinically justified in principle.

As an example of the first situation, consider a patient (P) with acute appendicitis who refuses a life-saving appendectomy. Table 1 presents some of the reasons CLAD knows about for refusing surgery, framed as **beliefs** held by P, as well as some **clues** in favor of each analysis.

CLAD may or may not have information about these clues stored in P^{C}_{i} (its model of the clinician's model of P_i). If it does have such clues, it can hypothesize about which belief is leading to the patient's decision. This hypothesis alone might be sufficient to help a tired, frazzled, rushed – in general, cognitively overloaded – clinician to steer the conversation with P in a useful direction. However, if the clinician wants more help, CLAD can suggest the best argumentation strategy by combining known methods to address the belief (see the 'How to…' column in Table 1) with known features of P that affect the choice (see the 'Influencing features…' column in Table 1). For example, if P is a Christian Scientist with low medical

sophistication but high intelligence, the clinician might engage him in a debate about the details of Christian Science, whose theology does not actually require refusal of medical intervention. If, by contrast, P is a Christian Scientist with a high fear of death, low intelligence and low medical sophistication, the clinician might better choose to focus on the statistical likelihood of dying – that is, if the clinician decides that it is ethically appropriate to try to change the patient's mind to begin with.

Belief	Clues	How to Address Belief?	Influencing Features of P
The body can heal itself.	P has said this. P doesn't have regular check-ups, vaccines, etc.	Statistical likelihood of dying. Physiological explanation.	P's medical sophistication. [Fig. 2] P's cultural background. P's level of fear.
Surgery is too dangerous.	P has said this. P's relative died in surgery. P has rejected indicated surgery in the past.	Statistical likelihood of dying. Break down aspects of surgery, discuss risks of each.	P's medical sophistication. P's level of fear. P's foci of fear.
Christian Science.	P has said this. P has never accepted a medical intervention. P has no medical history.	Statistical likelihood of dying. Physiological explanation. Primer on Christian Science theology.	P's medical sophistication. P's level of fear. P's foci of fear. P's intelligence.

Table 1. A subset of the appendectomy decision space.

A natural question is, how does CLAD acquire the beliefs stored in P^C_i? Some beliefs can derive from direct evidence: e.g., P's chart says he is a male, and CLAD assumes – with maximal confidence – that the clinician believes that. Other beliefs are derived through CLAD's reasoning. In the current version of CLAD, this reasoning is carried out using decision functions that take as input evidence from sources as varied as the patient's own statements, the patient chart, and the interpreted transcript of doctor-patient conversations. For example, the property MEDICAL-SOPHISTICATION is referred to in Table 1 as a feature influencing CLAD's decision making. Its value is calculated by CLAD using a stored decision function whose input parameters include EDUCATION-LEVEL, VOCABULARY-CHOICE, QUESTION-SOPHISTICATION and QUESTION-FREQUENCY. Each of these input parameters is assigned a time-dependent value, with CLAD assessing and recording its confidence in this assignment. Figure 2 illustrates the subset of P^C_i devoted to P's medical sophistication.

The last three types of evidence contributing to calculation of the value for MEDICAL-SOPHISTICATION all rely on functions that take as input the text-meaning representations (TMRs) of the doctor-patient interviews, which are generated by CLAD using the OntoSem language analysis system [13]. The decision function for VOCABULARY-CHOICE estimates the level of vocabulary use by the patient based on word length and comparisons with available dictionaries of difficult words; the function for QUESTION-SOPHISTICATION measures the content-oriented sophistication of the patient's questions based on how often they seek information about *how* or *why* some medical event happens, and how many words in questions are mapped to the medical subtree of the ontology; the value of QUESTION-FREQUENCY is a function of the average number of questions per visit, both direct (*Will it hurt?*) and indirect (*I suppose I won't need to take this medicine very long*). Naturally, estimating the sophistication of questions asked to the doctor is the most difficult, and therefore least confident, calculation.

HUMAN-301
- MEDICAL-SOPHISTICATION .8
  - CONFIDENCE .8
  - EVIDENCE
    - EDUCATION-LEVEL 1
      - CONFIDENCE 1
      - EVIDENCE INTAKE-QUESTIONNAIRE
    - VOCABULARY-CHOICE 1
      - CONFIDENCE .8
      - EVIDENCE TMR OF D-P INTERVIEWS
    - QUESTION-SOPHISTICATION .5
      - CONFIDENCE .5
      - EVIDENCE TMR OF D-P INTERVIEWS
    - QUESTION-FREQUENCY 10
      - CONFIDENCE .8
      - EVIDENCE TMR OF D-P INTERVIEWS

Figure 2. CLAD's belief about the doctor's belief about the medical sophistication of the patient.

Another way for the agent to infer property values of other agents is through static correlations among values of features comprising an agent model that we hypothesize might have predictive power. A sampling of such correlations is shown in Table 2. This table also gives a sample of the many kinds of features contained in CLAD's models of agents.

Assume that CLAD is faced with a decision whose function requires knowledge of an agent's level of optimism, and assume that, at the moment, CLAD does not have an explicit value for this property stored in its model of that agent. It can choose to estimate the agent's level of optimism (Col. 1) based on previous evidence of it being happy or depressed (Col. 2), making overly optimistic or pessimistic prognoses (Col. 3), and/or accepting challenges or avoiding risky behavior (Col. 4). CLAD's confidence in its estimation of the value of OPTIMISM depends on the amount of evidence available in its memory. An interesting question that we will not pursue in this short space is when (how often, at what junctures) it is appropriate for an agent to make generalizing conclusions about features of other agents like overall optimism, or having a high susceptibility to jumping to conclusions. Decisions like this are handled by the World Model Maintenance Engine shown in Figure 1.

Let us return now to the second class of CLAD functionality we use for illustration: CLAD helping the clinician to avoid incorrect prognoses. Making prognoses about things like the likelihood of a medication benefitting a particular patient is a tricky business. Clinical trials can provide information such as "medication efficacy: 50%." This means that for any given patient, this medication has an equal chance of being effective

and ineffective. If we consider that a positive attitude can positively impact healing, then a clinician might be justified in saying, "Of course it will work!" If, by contrast, we consider that offering false hope might cause a patient to lose trust in his doctor, a more coolly objective prognosis might be justified. The question then is: how can an intelligent assistant be useful to a clinician making prognoses? We suggest at least two ways. On the one hand, since our existing predictive physiological models were developed through an intensive effort by expert clinicians, they permit CLAD to make more specific prognoses than a clinician can be expected to make on the fly under the time pressure of an office visit (see [4,5]). On the other hand, CLAD can offer opinions about the extent to which overly optimistic or pessimistic decisions are justified based on parameter values found in its models of the patient and the clinician as well as in its knowledge about the objective medical situation. Let us consider a specific example of the latter in more detail.

Static Traits	Related Transient States	Related (Susceptibility to) Biases	Related Preferences
robust ↔ fragile	fresh ↔ tired	[none] ↔ depletion-effects & cognitive-overload	take-on-more-work ↔ avoid-more-work
optimistic ↔ pessimistic	happy ↔ depressed	overly-optimistic-prognosticating ↔ overly-pessimistic-prognosticating	accept-challenges ↔ avoid-risky-behavior
analytical ↔ impulsive	concentrated ↔ rushed	[none] ↔ jumping-to-conclusions & small-sample-bias	postpone ↔ act-now
confident ↔ insecure	decisive ↔ indecisive	illusion-of-validity ↔ cognitive-overload	convince-others ↔ let-others-decide-for-themselves
empathetic ↔ aloof	engaging-another ↔ keeping-distant	tendency-to-conceal-bad-news ↔ strictly-"like-me"-reasoning	close-relations ↔ distant-relations
extroverted ↔ introverted	chatty-mood ↔ terseness	[none]	long-conversations ↔ short-conversations

Table 2. Property correlations that help to fill out values of an agent model. ↔ indicates a scale whose end points are indicated. Features before and after ↔ correlate across columns: e.g., [optimistic, happy, overly-optimistic-prognosticating and encourage-others] are related.

In any given situation, CLAD can combine its general knowledge of medicine with known features of the patient to arrive at its objective prognosis for a medication's efficacy for the patient. As an initial simplification, CLAD has three available values for a medication's likely efficacy: *unlikely to work, might work, very likely to work*. Assume that the objective prognosis in a given situation is *might work*, and assume that the clinician tells the patient *Surely it will work!* Is this a misrepresentation (possibly a breach of ethics), a clinician error (e.g., due to an incorrect use of statistics) or a clinically justified decision on the part of the clinician ("My patient needs some hope and good news.")?

CLAD can attempt to trace the clinician's reasoning for presenting an overly optimistic hypothesis using a function that considers certain features of the clinician, the patient, and the clinician-patient relationship. If the reasoning seems justified, then CLAD will conclude that the exaggeration was intentional and will throw no warning.

The features of interest in this decision are shown in Table 3, which compares two different patients who have the same physiological profile but different character traits and different relationships with the doctor. The objective prognosis – *it might work* (based on "likelihood of treatment success for this patient: .5") – is always available and appropriate if the clinician chooses it. What CLAD needs to understand is whether an exaggeration like *Surely it will work!* is justified for either of these patients.

Feature	Source of value	Patient-A Value	Patient-B Value
Likelihood of treatment success at population level	function (literature, clinician's experience)	.5	.5
Likelihood of treatment success for this patient	function (population-level success, patient-specific features)	.5	.5
Best score of other available options	function, for each other available option (population-level success, patient-specific features)	.2	.2
Overall optimism of clinician	CLAD's past memories of clinician behavior	1	1
Confidence of clinician	as above	1	1
Clinician's knowledge about treatment	as above	.8	.8
Personal relationship between clinician and patient	as above	*1*	*.1*
Patient's medical sophistication	See Fig. 1	*.1*	*.8*
Patient's need for encouragement	as above	*1*	*.2*
Patient's susceptibility to encouragement.	as above	*1*	*.2*
APPROPRIATE PROGNOSIS: Likelihood of success of treatment is:	function (all of the above values and their confidences)	*.7 - .9 (Surely it will work!)*	*.5 (It might work.)*

Table 3. Sample calculations of clinically justified prognoses for 2 patients.

As the table shows, this prognosis is more easily justified for Patient-A than for Patient-B. Patient-A has low medical

sophistication (he is unlikely to know anything about the medication), a great need for encouragement and a high level of susceptibility to the doctor's encouragement; in addition, the doctor knows this patient well and feels justified in stretching the truth to fulfil these needs. By contrast, Patient-B is medically sophisticated and might know a lot about the medicine or look up that information later; and the doctor does not know him very well and does not sense any particular need for encouragement. As such, there is no clear justification for exaggerating the likely efficacy of the medication and it is better to report maximally objective prognosis. If a clinician should tell Patient-A that the medication will surely work, CLAD will interpret that as a clinically justified exaggeration, but if he should tell Patient-B the same thing, CLAD will throw a flag in case the clinician spoke in error. As mentioned above, CLAD's role is to assist the clinician in avoiding errors by trying to understand his reasoning; it is not CLAD's place to have an opinion about the judiciousness of exaggerating a prognosis to the positive.

3 INTERIM CONCLUSIONS

The development of an explanatory theory of mind that can be realized computationally in intelligent agents capable of being accepted as members of teams consisting of agents and humans (and not just efficiency tools like calculators or internet search engines) is arguably the most forward-looking ambitious research program in computer applications today. To be explanatory, this theory must, among other desiderata, account for reasons underlying agents' behavior. We believe that this is best done through the introduction of descriptive mental models that use parameters representing directly unobservable features of agents. This is a long-term project. But it is not too early to discuss how to model the moral stance of agents and what connections this stance has to the agent's theory of mind. Indeed, if such agents are to be accepted as team members by humans, then they will be expected to ne endowed with ethics – otherwise they would not be trusted by the human team members and would not be able to reason about other agents' motivations or explain their actions.

In this paper we illustrated the approach to the theory of mind in the OntoAgent environment, specifically concentrating on ethics-oriented features and situations. We discussed two levels of decisions – how to extract values for a variety of directly unobservable parameters and how to make decisions about a) other agents' beliefs and intentions and b) the agent's own actions on the basis of these parameter values. We extended the inventory of features (ontological properties) for modeling agents to include ethical considerations and so far have not found any problems with treating these features in exactly the same manner as other agent features.

In our work we followed the path established in [2] and concentrated on developing an advisor system, CLAD, that is constrained to clinical medicine. Unlike the Andersons, we also incorporated elements of mindreading in our agents, as a result of which CLAD's advising activity seeks to model human decision making in social environments where agents must model and take into account the "inner world" of other agents. This latter capability allows for modeling different ethical theories and different points of view within a single computational testbed.

A central task in the ongoing development of the OntoAgent theory of mind is enhancing the world model maintenance engine, including maintenance of ever more sophisticated models of self and other agents and the interaction between this task and language processing and other perception engines. A major component of this task boils down to knowledge acquisition for specific applications, and we intend to continue acquiring relevant knowledge by working with domain experts and by interpreting the findings of psychological experiments suggesting certain generalizations about human behavior. We will also continue our application-building efforts for the purposes of testing and validating the theoretical hypotheses.

At the same time, we will continue our work on formulating the theory of agent's mind as such. While at this point it is premature to offer a comprehensive description of this theory, we intend to formulate one in the near future.

REFERENCES

[1] Bello, P. 2011. Shared Representations of Belief and Their Effects on Action Selection: A Preliminary Computational Cognitive Model. Proceedings of the 33rd Annual Conference of the Cognitive Science Society.

[2] Anderson, M., S.L.Anderson and C. Armen. An Approach to Computing Ethics. 2006. IEEE Intelligent Systems, July-August.

[3] Ross, W.D. 1930. **The Right and the Good**. Clarendon Press.

[4] McShane, M., Fantry, G., Beale, S., Nirenburg, S. and Jarrell, B. 2007. Disease interaction in cognitive simulations for medical training. Proceedings of MODSIM World Conference, Medical Track, 2007, Virginia Beach, Sept. 11-13 2007.

[5] McShane, M., Nirenburg, S., Beale, S., Jarrell, B. and Fantry, G. 2007. Knowledge-based modeling and simulation of diseases with highly differentiated clinical manifestations. 11th Conference on Artificial Intelligence in Medicine (AIME 07), Amsterdam, The Netherlands, July 7-11, 2007.

[6] Nirenburg, S., McShane, M., Beale, S., English, J. and Catizone, R. 2010. Four kinds of learning in one agent-oriented environment. Proceedings of the First International Conference on Biologically Inspired Cognitive Architectures (BICA), Arlington, VA, Nov. 13-14.

[7] McShane, M., Nirenburg, S. and Beale, S.. 2008. Two Kinds of Paraphrase in Modeling Embodied Cognitive Agents. In Proceedings of the Workshop on Biologically Inspired Cognitive Architectures, AAAI 2008 Fall Symposium, Washington, D.C., Nov. 7-9.

[8] Nirenburg, S., McShane, M., and Beale, S. 2008. A Simulated Physiological/Cognitive "Double Agent". In Proceedings of the Workshop on Naturally Inspired Cognitive Architectures, AAAI 2008 Fall Symposium, Washington, D.C., Nov. 7-9.

[9] Nirenburg, S., McShane, M., and Beale, S. 2009. A unified ontological-semantic substrate for physiological simulation and cognitive modeling. In Proceedings of the International Conference on Biomedical Ontology, University at Buffalo, NY, July 24-26, 2009.

[10] McShane, M., Nirenburg, S. and Beale, S. 2011. Reference-related memory management in intelligent agents emulating

humans. Proceedings of AAAI Fall 2011 Symposium on Advances in Cognitive Systems.

[11] McShane, M., Jarrell, B., Fantry, G., Nirenburg, S., Beale, S. and Johnson, B. 2008. Revealing the conceptual substrate of biomedical cognitive models to the wider community. Medicine Meets Virtual Reality 16, ed. J. D. Westwood, R. S. Haluck, H. M. Hoffman, G. T. Mogel, R. Phillips, R. A. Robb, K. G. Vosburgh, 281 – 286.

[12] Nirenburg, S., McShane, M., and Beale, S. 2010. Aspects of metacognitive self-awareness in Maryland Virtual Patient. Proceedings of the AAAI Fall Symposium on Cognitive and Metacognitive Educational Systems, Nov. 11-13, Arlington, VA.

[13] McShane, M. and Nirenburg, S. 2009. Dialog Modeling Within Intelligent Agent Modeling. Proceedings of the IJCAI-09 Workshop on Knowledge and Reasoning in Practical Dialog Systems, Pasadena, California, USA, July 12, 2009, pp. 52-59.

[14] Nirenburg, S. and Raskin, V. 2004. Ontological Semantics. Cambridge, Mass.: The MIT Press.

[15] McShane, M., Nirenburg, S. and Beale, S. Ms. Meaning-Centered Language Processing. Book-length manuscript, submitted.

[16] McLaren, B. 2006. Computational Models of Ethical Reasoning: Challenges, Initial Steps and Future Directions. IEEE Intelligent Systems, July-August.

Machine Ethics, Folk Intuitions, and the Cognitive Architecture of Moral Judgments

Paul Bello[1] and **Selmer Bringsjord**[2]

Abstract. This paper begins the exploration of a new research paradigm for machine ethicists: a systematic focus on the mental representations and processes that produce commonsense moral judgments of the variety that all normally developed humans seem to be capable of. We assume that formally capturing the relevant conceptual repertoire along with developing properly parameterized inference mechanisms satisfy the necessary and sufficient conditions for building a machine equipped with something like robust moral commonsense. After discussing the various advantages and challenges of taking this particular tack on machine ethics, we explore a case study involving the interplay of intuitions about freedom, responsibility, and the self. Specifically, we examine recent results in experimental philosophy that provides a richer picture of the set of concepts involved in moral judgment, and speculate that some of the trends existing across the data are explicable in light of the cognitive architecture of mental state attribution or *mindreading*, as we shall refer to it. We suggest that along with machine ethicists working on the implementation of meta-ethical principles generated in the armchair, we ought to pursue the formalization of folk intuitions about freedom and agency to move us closer toward moral machines. So long as a robot has something like human folk beliefs about freedom and agency, and can deploy these believably in service of moral evaluation, it looks as if we might avoid the dispute about the correct (meta)ethic to adopt in favor of outright trickery: a fitting strategy for this celebration of Alan Turing's life and work.

1 Introduction

If we're ever to move past our flirtation with the idea of moral machines into a serious engagement, we must build the sort of system that we're both intellectually and emotionally willing to heap just desert upon. By this, we mean systems that behave in ways that compel us to strongly anthropomorphize in the way we sometimes do for cartoon characters, animals, and other non-human entities. But in this case, the consideration of robot-as-responsible-person must be more than just a fleeting ascription of agency, followed by a episode of convincing ourselves that the thing we're looking at is a robot devoid of an inner life. After all, there isn't much sense in sanctioning a machine that doesn't have desires, hopes and wishes; or who didn't sincerely believe that if given the opportunity, it could freely choose to pursue them and act to achieve them. This kind of system would have to have a sense of *self* that persists through time; an *I* whose future prospects for self-satisfaction might be limited in cases where it was judged blameworthy, and sanctioned by having its freedom limited. Building a machine that is capable of even the rudimentary forms of mental-state reasoning, causal attributions, and inferences about ability would presumably be a significant step in the right direction. Even being able to develop the requisite ontological and conceptual resources for building a simple moral agent would be a considerable advance. Unfortunately, many of the formal frameworks and computational tools on offer define a space of hyper-rational artificial agents. Such systems will slavishly obey a set of ethical constraints, but might well produce completely incomprehensible moral judgments, or take actions that seem out-of-step with commonsense moral intuitions. Just looking at the now-infamous *Trolley Problems*, what we find are human subjects having systematically different moral intuitions (consequentialist versus deontological), about what to do in near-identical moral dilemmas. If our moral machines are built by merely adopting some or other meta-ethical framework, it is likely to generate moral judgments that are seriously at odds with human moral commonsense.

Some might argue that there are some domains in which human moral commonsense has been appropriately refined into a set of implementable rules; in particular, we're thinking of the Laws of Armed Conflict [1]. We don't fundamentally disagree, but machines that have been built to optimize over formalized versions of such rules don't strike us as being genuinely moral. The machine is implicitly obligated to do its very best at choosing actions that satisfy the rules, but this is merely constraint satisfaction. There isn't really a sense that currently implemented systems understand obligation in a way that allows for intentionally flouting such constraints without good reason. Worse than this, the space of domains in which humans are capable of making moral judgments is more or less infinite. Success in a single domain with well-defined rules of the road ought to be lauded, but lingering questions remain about how such approaches generalize. Indeed, it has recently been shown that an ethical code based directly upon a divine-command ethics can be formalized, rendered computational, and implemented [3] — but by definition a robot constrained by such a code will be accepted only by those accept, for starters, that God exists, and that he has laid down commands constitutive of what is right and wrong.

It thus would seem to be a desideratum that our focus be on understanding the cognitive mechanisms that produce the highly varied, yet systematic patterns of moral judgments that we see across human beings. Our hope is that by moving further from learned sets of rules down toward the cognitive mechanisms that manipulate mental content during episodes of moral evaluation, we have a better chance at building a robust moral cognizer. What we're talking about here entails enormous amounts of (yet-to-be-done) research, formalization

[1] Office of Naval Research, 875 N. Randolph St., Arlington VA 22203 USA, email: paul.bello@navy.mil

[2] Depts. of Cognitive Science & Computer Science & the Lally School of Management & Technology, Rensselaer Polytechnic Institute (RPI), Troy NY 12180 USA, email: selmer@rpi.edu

and implementation that we can't even start to pursue in earnest in this document; but we think it's worthwhile to begin exploring the some of the conceptual materials that seem to be employed in human moral judgments. To do so, we turn briefly to some of what's been recently learned about the nature of folk concepts about freedom, agency, and responsibility by way of *experimental philosophy*. After pulling out some common themes from the human data, we attempt to explain some of the more interesting trends in judgment as a function of how *mindreading* might be operating. In fact, we hope to begin building a case for so-called *simulationist* approaches to mindreading based in part on our reading of these results.[3]

2 Intuitions about Freedom and Responsibility

The two central questions in the philosophical debate over free will is whether or not we indeed have it, and if so, whether or not it is compatible with determinism. Most philosophers agree that free will is closely linked to moral responsibility, and roughly define freely chosen acts as those which provide grounds for assignment of praise or blame. Enormous amounts of literature on these questions has been generated over the years, but all of it has assumed a level of theoretical sophistication and training that one is unlikely to find in an untutored member of the population. A growing group of philosophers have become interested in peoples' pre-theoretical evaluation of philosophical issues (such as the compatibility question) expressed in a manner that leaves the wiredrawn niceties of analytic philosophy to the side, and have brought the tools of empirical science to bear on studying these intuitions. Such studies have been typically called "experimental philosophy," and thankfully for us, many of them have focused exclusively on questions surrounding our intuitions about freedom, agency, and responsibility.

Eddy Nahamias and colleagues [9] devised a series of scenarios in which an agent performs a moral or immoral action in a deterministic universe. In the first experiment, reminiscent of the film *Minority Report*, subjects were told that scientists of the next century have discovered all relevant laws of nature and have programmed a supercomputer to be able to predict the course of all events with 100 percent accuracy. They were then told that the computer predicts that Jeremy will rob a bank at a particular time on a particular day. They are then told that true to form, Jeremy indeed robs the bank on the predicted day and time. Subjects were then asked whether Jeremy was morally blameworthy for robbing the bank. Similar scenarios were presented for morally neutral and morally praiseworthy actions. In all cases, subjects strongly tended to give compatibilist responses, claiming that Jeremy was indeed morally responsible for robbing the bank. Suspecting that the definition of determinism given to the subjects might have influenced responses, the experiment was repeated with new descriptions of determinism: one involving Jeremy's genetic/environmental predispositions, and one involving a universe that was recreated and started over with exactly the same laws and exactly the same events obtaining based on those laws. In both cases, Nahamias et al. discovered the same results; this led them to question the assertion that people are natural incompatibilists, and that compatibilism is a position only arrived at after digesting the carefully crafted arguments of philosophers.[4] Nichols & Knobe designed similar experiments, but varied the language used to describe the vignettes along an abstract-to-concrete continuum [10]. For example, after giving a fairly standard definition of determinism (and indeterminism), subjects are asked whether an agent can be morally responsible in a determined universe (the abstract condition). The majority of subjects responded in an incompatibilist way, saying "No." Subjects were then told about Bill, who in the same deterministic universe burns down his house with his wife and children inside so he can run off with his secretary (the concrete, affect-laden condition). Subjects overwhelmingly responded in a compatibilist way, saying that Bill was indeed morally responsible. In a third example, subjects are told about a serial tax evader, and asked whether or not he was morally responsible for tax evasion (the concrete, low-affect condition). In this case, subjects tended to not blame the tax evader — certainly not to the degree that they blamed Bill the murderer. Similar scenarios were presented in a universe described as indeterministic, with subjects responding in the expected way: blaming the responsible parties in all cases. What the Nichols-&-Knobe results suggest is that there are outside influences on our attributions of responsibility that need to be explained, and certainly need to be accounted for in any computational model that seeks to reproduce human performance on moral judgment tasks.

2.1 Intuitions about the Self

More recently, Nichols and Knobe ran a complimentary study on people's intuitions about the nature of the self as it pertains to responsibility [8]. These studies specifically avoided the compatibilism/incompatibilism question in favor of probing intuitions about the nature of the folks' incompatibilist intuitions: why they are reluctant to blame in the abstract, low-affect cases presented in [10]. Specifically, different philosophical conceptions of the self were studied: one in which the self is identified with the body, one in which is self is identified with psychological states,[5] and one in which the self is identified with a "central executive" above and beyond the self-as-body or self-as-psychological-state. Nichols and Knobe contend that there is a core reason why these three conceptions of self have been studied so intensely from a philosophical perspective: namely, because all three are used in making judgments about how the self relates to action under varying circumstances. In particular, they propose that given some agent A, people will deploy a bodily/psychological notion of self-as-cause when A's actions are considered in a broader situational context. On the contrary, when we zoom in to look at the action itself and the mental processes surrounding it, people will tend to deploy the executive notion of self, treating A itself separately from the processes surrounding the action. For further clarity, we give a description of some of the stimuli, and here quote directly from the source laconic materials, which decidedly leave behind the longuers seen in the thought-experiments of professional philosophers:

> Subjects were randomly assigned either to one of two conditions. In one condition, subjects received what we will call the choice-cause case:
>
> Suppose John's eye blinks rapidly because he wants to send a signal to a friend across the room. Please tell us whether you agree or disagree with the following statement:

[3] Such approaches should not be here understood to preclude modeling cognition via formal logic. On the contrary, simulation of cognition is if anything wisely pursued on the strength of what logic has to offer. E.g., see [2]; and for a general account of logic-based cognitive modeling, see [5].

[4] For the record, in the face of results apparently showing that logically and philosophically untrained subjects often lean toward compatibilism, Bringsjord is a staunch incompatibilist, and stands by his original argument for both incompatibilism and agent causation, presented in [4].

[5] This isn't a bad spot to point out that neither of us, in citing Nichols and Knobe, indicates thereby an affirmation of the ontological presuppositions they appear to make. For instance, Bringsjord is rather convinced that as such notion that the self is but a collection states is in the end incoherent. The self for him is the thing that is the bearer of psychological states.

- John caused his eye to blink.

In the other condition, subjects received what we will call the *emotion-cause* case:

Suppose John's eye blinks rapidly because he is so startled and upset. Please tell us whether you agree or disagree with the following statement:

- John caused his eye to blink.

Subjects rated each statement on a scale from 1 ('disagree') to 7 ('agree').

As predicted, subjects generally identified John as the cause of his eye-blinking in the choice-cause condition, while asserting that John wasn't the cause of his eye blinking in the emotion-cause condition. Consistent with their "zooming" account, the zoomed-in description of the mental circumstances surrounding John's eye-blink in the emotion-cause condition compelled subjects to deploy the John-as-executive conception of self, whereas in the zoomed-out choice-cause condition, subjects deployed the John-as-psychological-states conception of self. A second experiment was run to rule out the possibility that "ordinary folk" don't consider being startled as the kind of psychological state that's constitutive of persons. Subjects were told that "John's hand trembled because he thought about asking his boss for a promotion." They were then asked to agree (on a 1–7 scale) with the contrasting statements: (1) *John caused his hand to tremble*, and (2) *John's thoughts caused his hand to tremble*. Consistent with results from the first study, people tended to agree with (2) much more than (1). In a third condition, subjects were given the following:

Suppose that John has a disease in the nerves of his arm. He experiences a sudden spasm, his arm twitches, and his hand ends up pushing a glass off the table. As the glass strikes the floor, there is a loud crashing noise.

Then, the subjects were given two questions, a "zoomed-in" question, asking them to agree or disagree with the statement "John caused his arm to twitch," and a "zoomed-out" condition asking them to agree with the statement "John caused the loud noise." Again, the results showed subjects willing to agree with the assertion that John caused the loud noise, but disagree with the assertion that John caused his arm to twitch. The pattern of responses given suggests that by asking questions that "zoom out" and consider the situation more broadly, our intuitions lead us to adopt the John-as-body (similar to John-as-psychological-state) notion in our causal attributions. To proceed *ad rem*, we won't review the last of the experiments, which varied both zooming and the type of action (choice-cause vs. emotion-cause); but the results are predictable, and consistent with the multiple-self-concept hypothesis.

2.2 Summary of Experimental Results

In our very brief tour through some of the most recent literature on folk intuitions, we saw clearly distinct patterns of judgment. When it comes to free will, the jury still seems to be out on whether or not humans are natural compatibilists or natural incompatibilists.[6] The data suggest that how abstractly or concretely the decision-problem is framed makes a big difference to our judgments about responsibility. Similarly, the degree to which a scenario is affectively valenced seems to also make a difference. The Nichols-&-Knobe study on the nature of self-concept(s) suggests that a sufficiently "zoomed out" description of an agent's activity leads to the agent being held responsible for outcomes. This is in contrast to when we consider the immediate mental circumstances the agent finds itself in prior to the act, in which case we're less likely to ascribe responsibility to the agent, especially if the acts don't seem to cohere with the agent's putative desires in the right kind of way. We saw this in the case of John's twitchy arms, eyes, and hands. It looks increasingly likely that untrained intuitions about freedom, agency, and responsibility are driven by other factors, including the perspectival nature of how moral decision problems are framed. But how might we explain these trends in the data without resorting to positing multiple conceptions of self, or having fairly elaborate definitions of causality on hand?

[6] By use of the adjective *natural* we mean to reference untutored human intuitions. The subjects in the experiments weren't professional or graduate-level philosophers; this presumably keeps any sort of bias introduced by prior exposure to the literature on free will to an absolute minimum.

3 Mindreading and Moral Cognition

Mindreading fundamentally concerns the human ability to predict and explain the behavior of other agents in terms of their beliefs, desires, intentions, and other mental states. Theories abound when it comes to what the cognitive architecture of mindreading looks like. We will focus on a particular variety of theory called *simulation theory* (ST). There are different versions of ST [6, 7], but all fundamentally hold that mindreading is largely composed of a mechanism that allows the mindreader to "step into the mental shoes" of the target agent, using its own inferential resources as a first-pass approximation for the target's. The simulation strategy clearly obviates the need for representing entire theories of practical reasoning as being in the possession of other agents. The most well-developed simulation theory construes mindreading as populating a mental simulation with pretend beliefs, desires and other mental states, followed by an inferential elaboration stage in which the simulator's inferential resources are brought to bear on the contents of the simulation. Finally, the results of the simulation are "taken off-line," so that the simulator doesn't end up with the results of these simulations having any effect on its own cognitive state. Even with the profusion of hypotheses about the nature of mindreading, many researchers are starting to see a role for mental simulation, especially when attempting to mindread a target about whose mental states you are somewhat uncertain.

However, if humans happen to be natural incompatibilists, then mindreading becomes somewhat more complicated. As natural incompatibilists, humans would have requisite beliefs about agent-causal freedom that undoubtedly get ascribed during episodes of mindreading. To be clearer, when agent A attempts to predict agent B's behavior, A will ascribe some of its relevant set of beliefs to B in order to do so. If A believes that agents can (generally) do otherwise than what they are currently intending, and via simulation believes that B believes this, A is stuck trying to make predictions about an agent who believes it can do otherwise. This, of course, is quite an inhospitable set of circumstances for effective mindreading. It seems like reasoning about all of the things an agent could have otherwise done in similar circumstances only serves to make prediction more difficult, and explanation more circuitous.

The first question we're prompted by the Nichols-&-Knobe results to ask is: Why don't we generally blame agents in situations that aren't richly described? Naturally, the second question is what moves us to blame them when we add more information to the description of the situation. As for the first question, we think the answer might be buried in the mechanism by which we run mental simulations of both self and others. The mental simulation process almost always involves some degree of counterfactual reasoning. Mental simulations used for prediction of behavior naturally involve

the present-tense counterfactual conditional: "If I were x in situation y then I'd do/think/hope... z." Simulations used for explanation of behavior use the more commonplace past-tense counterfactual conditional: "If I were x and executed (mental or physical) action y then I would have thought/hoped/intended... z." Insofar as our intuitions about libertarian freedom involve inferences of the form "If I thought/hoped/wished... a, I could do/think/hope... b," they look to be implemented by the same mental simulation process that we deploy when predicting or explaining the behavior of others. In cases where we might imagine ourselves with different mental states, it seems that the connection between acting and having a particular set of mental states is a tenuous one. We hypothesize that having the ability to simulate different versions of oneself being in various states of mind leads to very weak priors on the conditional probability of a particular action, given a particular set of mental states. On this account, we are naturally biased against (some of) our mental states[7] being causal factors in our actions in the absence of extra information. In these cases, there are limited options to tag as strong causal influences on action. We suspect that these simulations merely "fail" due to lack of high-confidence inferences made within the simulation. In the case where we're only told about John's twitch, few if any further inferences are invited. No descriptions of changes in the environment are described, as opposed to the case of the loud noise. Given a lack of situational anchor or expressed desire on which to pin an explanation, the simulation fails to generate an adequate explanation, and John remains unidentified as the cause of the twitch. Once we introduce very strong desires or highly salient situational constraints into our counterfactual inferences about our alter-egos, we might find libertarian intuitions much harder to hold onto.

The answer to the second question involves adverting to situational constraints in order to narrow the space of possible predictions and/or explanations generated during mindreading. The more we can say about a situation, the easier it seems to take a guess at what will happen next, or to explain why an observed outcome happened. Situational constraints provide anchors on which to hang explanations or to generate predictions. Our admittedly speculative hypothesis is that when we makes judgments about agents and causes, as in the Nichols-&-Knobe experiments about the self, we mentally simulate ourselves as the protagonist agent, and populate the simulation to the extent that the problem description and our prior knowledge allows for. We equate "zoomed out" cases in Nichols-&-Knobe parlance to mental simulations of us-as-protagonist being relatively more filled out with situational constraints and explicit information about desires than other "less zoomed out" simulations. We suspect that outcome-related information, such as in the case of the loud noise produced by the falling glass in Nichols-&-Knobe's stimuli, invites further inferences which require explanation. In the case of the noise, we begin to reason backward to the glass falling due to John's arm making contact, and back further to the twitch. Once involved in generating this sort of causal explanation, we assume it's likely to be the case that a global explanation is sought out to unify the sequence of mini causal-chains, leading toward a greater number of agent-related causal attributions. We suspect that in the case of John causing the loud noise, the inference to John's arm hitting the glass might sway judgment of responsibility closer to John than would be the case if we were just told about John experiencing a twitch. We've probably developed pretty strong priors on our body parts having causal import to changing the state of the world — priors that might be much more stable than those associated with mental states having similar efficacy. All of this remains to be seen, but what seems clear is that efficient operation of mindreading in either a predictive or explanatory capacity requires a sufficient degree of concreteness in the mental simulations that drive it. Similarly, the more concrete information we have about a particular agent-related event in the world, the more likely we are to assign responsibility to agents *qua* agents. The less information we have about an agent's circumstances, the more guesswork we need to do in simulation in order to produce an explanation. We gather that many of these simulations fail due to relatively weak priors on the relationship between mental states (again, we qualify this in footnote 7) and individual actions.

[7] We assume that this might not hold universally. It could be the case that certain kinds of mental states are harder to ascribe to oneself when they diverge from one's current mental states. Imagining a version of yourself as full when you're actually hungry is probably much more difficult than thinking you're the King of France when you're actually not. The fairly strong connection between desires and acting might be an exception to the general rule that people naturally don't give much weight to mental states being hard and fast constraints on action. We come back to this point at the end of the paper.

4 Summary and Future Research

If our examination of the human data on responsibility ascription tells us anything, it's that we've got a lot of work to do in order to capture the variance on display when faced with everyday instances of agent-causal judgments. On the supposition that we're right about the cognitive architecture of mental simulation, it might very well be possible to explain a substantial portion of this variance without resorting to having different concepts of self, or relying on complex folk-physical theories about determinism. As our ideas developed, it struck us that the Nichols-&-Knobe studies almost invariably resulted in responsibility ascriptions to agents-in-themselves whenever said agents had an expressed desire to perform certain actions (like John blinking to signal his friend). It was only in cases where no overt desires were expressed that variance in judgments was observed. Further studies might be done to control for the presence or absence of an agent's desire, and the linguistic expression of its relative strength. Furthermore, many of the actions described in the vignettes were involuntary, even the thoughts that led to John's trembling hand. This might also be an interesting variable to manipulate in follow-up studies. For our part, we are working toward a computational treatment of the Nichols-&-Knobe studies using an implemented theory of simulationist mindreading developed by the first author. We expect that the exercise of implementation will lead to further questions, more research, and eventually a richer computational story about moral cognition.

REFERENCES

[1] R. Arkin, *Governing Lethal Behavior in Autonomous Robots*, Chapman and Hall/CRC, 2009.

[2] K. Arkoudas and S. Bringsjord, 'Propositional Attitudes and Causation', *International Journal of Software and Informatics*, **3**(1), 47–65, (2009).

[3] S. Bringsjord and J. Taylor, 'The Divine-Command Approach to Robot Ethics', in *Robot Ethics: The Ethical and Social Implications of Robotics*, eds., P. Lin, G. Bekey, and K. Abney, 85–108, MIT Press, Cambridge, MA, (2012).

[4] Selmer Bringsjord, 'Free Will', in *What Robots Can and Can't Be*, 266–327, Kluwer, Dordrecht, The Netherlands, (1992).

[5] Selmer Bringsjord, 'Declarative/Logic-Based Cognitive Modeling', in *The Handbook of Computational Psychology*, ed., Ron Sun, 127–169, Cambridge University Press, Cambridge, UK, (2008).

[6] A. Goldman, *Simulating Minds:*, Oxford University Press, 2006.

[7] P. Harris, *The Work of the Imagination*, Blackwell Publishers Ltd., 2000.

[8] J. Knobe and S. Nichols, 'Free will and the bounds of the self', in *Oxford Handbook of Free Will: Second Edition*, ed., R. Kane, 530–554, Oxford University Press, (2011).
[9] E. Nahmias, S. Morris, T. Nadelhoffer, and J. Turner, 'Is incompatibilism intuitive?', in *Experimental Philosophy*, eds., J. Knobe and S. Nichols, 81–104, Oxford, (2008).
[10] S. Nichols and J. Knobe, 'Moral responsibility and determinism: the cognitive science of folk intuitions', in *Experimental Philosophy*, eds., J. Knobe and Nichols, 105–128, Oxford, (2008).

Machine Ethics, Mindreading and Attributions of Responsibility : First Computational Steps

Paul Bello[1] and **Selmer Bringsjord**[2]

Abstract. In a sister paper submitted to this symposium [4], we explored interesting new data generated by experimental philosophers on human attributions of responsibility [10]. This data suggests that human decision-makers deploy multiple concepts of "self" in support of causal attributions. Upon investigating trends in the data, we hypothesize that a significant portion of the variance might be due to the cognitive architecture of the human capacity to *mindread*. By mindreading, we refer to the human ability to predict and explain the behavior of agents by representing and reasoning about their mental states and inferential tendencies. In the present paper, we build on a pre-existing computational model of mindreading [3], showing how the variance in the aforementioned data on causal attributions might well be related to the set of architectural assumptions required to make mindreading tractable.

1 Introduction

As argued in [4], machine ethicists hoping to build artificial moral agents would be well-served by taking heed of the data being generated by cognitive scientists and experimental philosophers on the nature of human moral judgments. In prior work [6], we advocate for the building of artificial moral agents (AMA) that make *provably correct* moral judgments on the basis of formal reasoning using suitably rich logical formalisms. It appears that other machine ethicists have similar intentions. Ron Arkin's "Ethical Governor" is a prime example of just such a system [1]. Yet, however noble these goals, it's clear to us now that both Arkin's work and our prior work share a rather glaring omission in common. If our goal is to build an AMA whose moral capacity somehow exceeds that of a typical human, it seems reasonable to assume that at some point other human agents and their capacity for moral judgment will be the objects of our AMA's superior moral reasoning faculties. Of course, this almost necessitates being able to computationally reproduce the variance in prototypical human moral judgments. So regardless of whether you might be a machine ethicist looking to code up the "right" meta-ethic for your system, or whether you're a computational cognitive modeler interested in the nature of human moral judgments *sans* normativity, the need to account for the latter in our own attempts at building AMA's will invariably arise. Once we've committed ourselves to addressing this requirement, we need to start to determine the conceptual content and inference mechanisms supporting human moral judgment, and whether or not we can adequately approximate them using the tools we have at our disposal.

[1] Office of Naval Research, 875 N. Randolph St., Arlington VA 22203 USA, email: paul.bello@navy.mil

[2] Depts. of Cognitive Science & Computer Science & the Lally School of Management & Technology, Rensselaer Polytechnic Institute (RPI), Troy NY 12180 USA, email: selmer@rpi.edu

Being committed to modeling moral judgment through understanding and computationally capturing the relevant psychological constructs has the added benefit of providing a degree of generality that other approaches might not offer. Many of the implementations that exist in the machine-ethics literature are tied deeply to domain-specific rules generated by trained ethicists. While laws governing armed combat, or those defining criminal behavior, may represent something like a set of highly distilled human intuitions, they clearly don't completely cover the space of all human moral judgments. They certainly don't provide anything like a complete decision-procedure for assigning blame in the real world, which is full of resource-bounded reasoners. Approaches to case-based moral reasoning [9, 11] might provide computational avenues toward extracting common structure from disparate moral domains. In theory, common structure might reflect the operation of the principal components of moral cognition. Once we have these in hand, there might be some way to generalize them to other unseen domains; but to date, we don't see any viable computational ways to do so.

In any case, every computational approach to date helps itself to at least a few assumptions about the content it computes over, and the means by which such content is manipulated. Beginning to identify the content and mechanisms is the task to which we now turn, starting with the concept of *self* employed in everyday judgments about responsibility. We first provide a brief summary of some recent work on human intuitions about the self and causation, followed by a recapitulation of the hypothesis we presented in [4] regarding mindreading and its relationship to causal attributions. We then outline a computational model of mindreading, and use it to instantiate a working version of our hypothesis, showing that it does indeed reproduce general trends in the human data. We wrap up with a general discussion of our results and with some directions for the future.

2 Intuitions about Freedom and Responsibility: A Summary

The model we present below is of data on folk intuitions about the self and responsibility generated by the experimental philosophers Shaun Nichols and Joshua Knobe (henceforth N & K). Considered pre-philosophically, it seems likely that we would endorse the concept of a single "self" that is responsible for its actions by way of other agentic concepts like freedom to choose, fealty to obligations, and so on. But this is not exactly what the data reveals. Let's move further into the details by way of some of the examples given to subjects in the experiments N & K report on in their target paper [10].

In prior work, N & K show that attributions of responsibility are apparently related to how moral decision problems are *construed* by the reasoner. By construal, we largely mean the level of detail at

which the problem is described. In an attempt to investigate whether or not the untutored subject has natural leanings toward moral judgments being compatible with determinism, N & K describe for subjects part of what it would mean to live in a deterministic universe, and then present several vignettes against this backdrop of determinism. They first ask the rather abstractly construed question: "Is moral responsibility possible in a deterministic universe?" The majority of subjects responded with "No," as if they had incompatibilist leanings. However, when presented with an affectively charged, concretely described scenario such as an agent killing his wife and children in order to run away with his secretary, subjects largely held the agent morally responsible for the deaths, even against the same deterministic backdrop. Given a less affectively valenced but similarly concrete example in the same deterministic universe, subjects were less likely to assign responsibility, leaving N & K suggesting that both the level of construal and the expected level of elicited affect seem to account for variance in human judgments of responsibility [13].

N & K's target paper attempted to further investigate why subjects are reluctant to blame in low-affect, abstractly construed cases. To similarly untutored subjects, N & K present vignettes that vary the "distance" between a described behavior and the agent-as-cause. To be clearer, N & K propose that different notions of self are at play in attributions of responsibility depending on how the circumstances surrounding the target behavior are described, much in the way that so-called "framing effects" influence choice under uncertainty [14]. As an example, N & K give subjects a vignette about John, who has a neurological disorder that causes him to occasionally experience twitches in his arm. John experiences a twitch and his arm knocks into a glass, which falls off the table upon which it sits and crashes to the floor. The subjects are then asked whether John is responsible for the twitch. In line with expectations, most subjects respond that John isn't responsible; but then the subjects are asked whether or not John is responsible for the loud noise. This time, rather surprisingly, subjects generally held John responsible for the loud noise. N & K hypothesize that the degree to which situations are described as "zoomed out" compel subjects to treat the agent as a causal factor more often than when situations are described in a "zoomed-in" way. By "zoomed-in," we mean descriptions that put very little causal distance between the agent and the observed effect, as in the case of John's twitching in the first question. N & K suggest that a conception of self-as-central-executive which governs mental states and their relation to action is employed when judging scenarios that are described using zoomed-in language. As descriptions are zoomed-out, N & K point out that a conception of self-as-body is being utilized in responsibility ascriptions. In these situations, John is functionally considered to be like any other event in the causal chain.

2.1 Our Hypothesis: Mindreading Meets Zooming

We find the bi- or (tri-)partite conception of self posited by N & K to be illuminating, but still somewhat unsatisfying. Why should we have two or three conceptions of self in the first place? What sort of evolutionary or social pressures selected for such an odd cognitive feature? In our sister paper, we argued that the cognitive architecture of the human mindreading capacity might take us some way toward grounding this rich notion of self uncovered in [10]. In particular, we explored how on a particular account of mindreading driven by mentally simulating oneself-as-the-target during the mental state ascription process might explain both the source of the self-as-executive and self-as-body conceptions that are central to N & K's analysis. We noted that the kind of incompatibilist intuitions associated with the self-as-executive might be the result of self-simulations or the imagining of oneself in different (mental) circumstances. It's seemingly easy to imagine oneself with different beliefs, desires, intentions. It seems rather likely that self-simulations of this variety are the source of our intuition that agents can (often) do other than what they currently intend. What follows from the could-do-otherwise intuition are fairly weak conditional relationships between being in any particular mental state and taking an associated action. As we mentioned in [4], the could-do-otherwise intuition is a downright disaster for mindreading. The goal of mindreading is to assign mental states to other agents in order to facilitate predictions and explanations. If the mindreader comes equipped with the could-do-otherwise intuition as we suggest above, and assigns it via simulation to his target, he will be placed in the unfortunate position of resolving the uncertainties it introduces when attempting to predict or explain the target's behavior. If mindreader-as-target has the belief "I can do otherwise," extra simulations need to be run that ground out available options for "doing otherwise."

However, if simulations are initially populated with situational constraints or with content that invites further inferential elaboration, the uncertainties associated with the could-do-otherwise belief might be effectively tamed. In this case, we use situational constraints to mean observed or inferred causal relationships between objects in the vignette under consideration. The more elaborated and constrained a simulation is, the more useful it will likely be in facilitating predictions or explanations. The situational information provide anchors on which to hang explanations and make predictions. We assume fairly strong priors on causal interactions between objects, even when some of those objects are closely related to the agent (such as his limbs, etc.). As more causal inferences are made within the simulation, the likelihood of chaining backward to one of these agent-related causal instruments is increased, leading to the kinds of attributions we see in the zoomed-out vignettes described by N & K. Situational constraints also work to nail down the set of mental states we ascribe to the target. When we as mindreaders imagine being in a familiar concrete situation, it seems reasonable that we also imagine having a fixed set of associated mental states. Once these are fixed during an episode of mindreading, prediction becomes possible because the causal chain originating from the requisite mental states leading all the way to the outcome in question becomes completed.

This is clearly not the final form of a solution, and perhaps it isn't even close. We haven't taken into consideration a typical mindreader's beliefs about agency and control, which would be relevant to making inferences about twitches and the like. That being said, we think that this explanation moves us a little closer to a motivated account of why we seemingly have multiple conceptions of self when making attributions of responsibility. But our job isn't done yet. As machine ethicists, we'd like to take this hypothesis all the way to a computational instantiation. Doing so ensure us that our ideas have enough structural integrity that they can be appropriately formalized. As an attractive side-benefit, we end up building a computational foundation for conducting further research as new data become available.

3 A Computational Model of Mindreading

What follows is a very brief description of a computational theory of mindreading that has been used to model early mental-state attribution in infancy [3], and errors in attribution [2]; and has been used to detail differences and similarities between mindreading and introspection [5]. The need for brevity precludes the possibility of provid-

ing a detailed defense of the model, so we will have to be satisfied with but an outline of the very basic set of underlying assumptions and computations. The task-model of the data in [10] that we present could be implemented using a variety of formalisms. We present the model as a set of weighted constraints in a logical language existing in the space between first-order logic with identity and second-order logic.[3]

3.1 Representation and Inference

Given the nature of our hypothesis, it shouldn't be surprising that we endorse a broadly simulationist approach to mindreading. In classic presentations of simulationism, it's often the case that the mindreader creates a series of pretend beliefs, desires and intentions, "running" these within a mental simulation of the target in order to produce a prediction or explanation. The mindreader operates over this pretend mental content using his own practical reasoning system as a rough-and-ready substitute for the target's inferential capabilities. The result of such simulations are "taken off-line" so that actions performed by the simulated target don't affect the current set of motor intentions held by the mindreader [8]. On our account, simulations of this kind are a particular kind of counterfactual reasoning in which the mindreader identifies with the target within a simulated state of affairs. Information that the mindreader knows about the real world is available within these mental simulations through a process called *inheritance*, which we explore in some detail in the next section. On simulation theories, mindreading involves entertaining a counterfactual statement of the form: "if I were him/her, I'd ϕ."

Representing and reasoning about counterfactuals involves keeping representations of real situations separate from representations of counterfactual situations. This being said, we embark on some formal preliminaries that detail a situation-centric representation that we will use throughout the rest of the discussion.

3.1.1 Knowledge Representation

An *atom* is a relation over one or more entities that takes a truth-value at a specific time in a situation (or *world*) as we will refer to them). In general, atoms are of the form *RelName*(e_1, e_2, . . ., e_n, t, w). The penultimate argument represents a temporal interval. We use the letter "E" to designate the temporal interval representing "at all times." The last argument defines the world in which the relation holds. We use the letter "R" to represent the agent's beliefs about reality (rather than about imagined or counterfactual worlds). We might therefore represent "Paul is hungry at noon." as IsHungry(paul, noon, R). To represent the converse, we use standard negation: ¬IsHungry(paul, noon, R). Arguments of the form $?e_i$ as in IsHungry(?agent, ?t, ?w) are unbound variables. Relation names can also be prepended with ?, allowing for quantification over relations. Constraints express contingencies between atoms. The standard logical operators $\wedge$ and $\rightarrow$ are used to construct constraints. All constraints are implicitly universally quantified. For example, IsHungry(?agent, ?time1, ?w) $\wedge$ LineOfSight(?agent, ?food, ?time1, ?w) $\rightarrow$ ReachFor(?agent, ?food, ?time2, ?w) expresses that if an agent is hungry at time1 and has line of sight on some food, then the agent will reach for the food at time2. Existentials can be introduced by having unbound variables on one side of a constraint that do not appear in the other side. To represent $\forall_x \exists_y (P(x) \rightarrow Q(y))$, we write P(?x, E, ?w) $\rightarrow$ Q(?y, E, ?w). In this case, the variable ?y doesn't appear in the antecedent, and thus acquires existential force. Finally and importantly, we are able to represent *soft constraints* that generate costs on the worlds in which they are broken. To write "All professors are usually nutty," we say Professor(?x, E, ?w) (.75)¿ Nutty(?x, E, ?w). What this constraint essentially means is that for any professor in any worlds at any time, they are very likely to be nutty. If we find a world in which there is a professor who turns out to not be nutty, that world incurs a cost of 0.75. Constraints written using the $\rightarrow$ conditional incur infinite cost if broken, and are called *hard constraints*. An explanation of inference using weighted constraints is beyond the scope of this paper, however details concerning the inference procedure we use in modeling the task in [10] can be found in [7]. In general, when soft constraints involving atoms indexed by some world w are broken, w is penalized by the cost associated with breaking the constraint. The inference process continues in this manner till a so-called "best world" is found in which hard constraints are maximally satisfied and costs associated with breaking soft constraints are minimized. The best world will consist of the set of atoms having w as a world argument, along with their respective truth values.

3.2 Simulations, Worlds and Inheritance

As mentioned, our simulation-based theories of mindreading rely centrally on the notion of entertaining counterfactuals. In order to stay on track, we avoid further motivation of the use of counterfactual reasoning as a substrate within which to run the mental simulations associated with mindreading. Instead, we focus on the notion of *inheritance* between worlds. Inheritance as it relates to mindreading can be thought of as the mechanism used to populate mental simulations. Information available to the mindreader become available in the counterfactual world where the mindreader is the same as the target through the inheritance process. In essence, inheritance defines the relationship between the world as the mindreader sees it, and the world as the mindreader thinks the target sees it. The most basic form of an inheritance rule is given below, and captures so-called "default ascriptions" of the form "if it's true for the mindreader, then it's true for the mindreader-as-target."

Def (1) ?Relation(?e_1, . . ., ?t, R) $\wedge$ IsCounterfactualWorld(?w, E, R) (*cost*)¿ ?Relation(?e_1, . . ., ?t, ?w)

Where *cost* takes a value in the range (0,1). Every time this constraint is broken because the target is ascribed ¬?Relation(?e_1, . . ., ?t, ?w) by assumption or via inference in *w*, costs are incurred. Given what formal machinery we have, we now move on to providing an example of this formalism at work on one of the N & K examples discussed in prior sections.

4 Accounting for the Data

We spend this section exploring the the vignette in section 2 regarding John and the twitch that knocks the glass off of the table. Given the formal apparatus presented in the last section, we can begin to construct a simple domain theory. We first write down constraints that roughly serve the purpose of being circumscriptive axioms [12] that minimize the number of event occurrences and causal relationships that hold in individual worlds:

[3] This is perhaps the spot to say that the present paper is devoted to investigating and suggestively modeling mindreading, not the formal niceties of Polyscheme. We hence spend no time on the issue of exactly how expressive Polyscheme is relative to standard markers like first-order logic and second-order logic, or where Polyscheme stands in relation to conditional logics (which would seem to the the class of logics most relevant to our modeling objectives herein).

c_1: IsA(?world, World, E, ?w) ∧ IsA(?ev, Event, E, ?w) ∧ IsA(?ag, Agent, E, ?w) (.10)¿ -Causes(?ag, ?ev, ?world, E, ?w)
c_2: IsA(?world, World, E, ?w) ∧ IsA(?ev, Event, E, ?w) ∧ IsA(?ce, Event, E, ?w) (.10)¿ -Causes(?ce, ?ev, ?world, E, ?w)
c_3: IsA(?world, World, E, ?w) ∧ IsA(?ev, Event, E, ?w) (.10)¿ -Occurs(?ev, ?world, E, ?w)

These constraints serve to minimize thinking about causal relationships or events during mental simulation unless we have on very good evidence that they actually obtain. We continue by expressing two more constraints that define causal chains for events. In short, the first constraint states that if one thing is caused by another, the latter is in the former's causal chain. The second constraint states that if a causal chain exists for an event, and something is known to cause the most distal event in the chain, then the former gets added to the chain and becomes the newest distal cause:

c_4: Causes(?e0, ?e1, ?world, E, ?w) ==¿ InCausalChain(?e0, ?e1, ?world, E, ?w)
c_5: Causes(?e0, ?e1, ?world, E, ?w) ∧ InCausalChain(?e1, ?e2, ?world, E, ?w) ==¿ InCausalChain(?e0, ?e2, ?world, E, ?w)

Next, we have some very simple causal relationships encoded about potential events described by the vignette. The right-hand side of these constraints mark the caused event as a new focal event. This will become important momentarily.

c_6: Occurs(loudNoise, ?world, E, ?w) ==¿ Occurs(glassFall, ?world, E, ?w) ∧ Causes(loudNoise, glassFall, ?world, E, ?w) ∧ FocalEvent(glassFall, ?world, E, ?w)
c_7: Occurs(glassFall, ?world, E, ?w) ==¿ Occurs(armMotion, ?world, E, ?w) ∧ Causes(glassFall, armMotion, ?world, E, ?w) ∧ FocalEvent(armMotion, ?world, E, ?w)
c_8: Occurs(armMotion, ?world, E, ?w) ==¿ Occurs(twitch, ?world, E, ?w) ∧ Causes(armMotion, twitch, ?world, E, ?w) ∧ FocalEvent(twitch, ?world, E, ?w)

The critical constraint for mindreading is given below. It captures the basic structure of trivial belief ascription by simulation, and implements the inheritance schema presented as Def (1):

c_9: IsA(?parentworld, World, E, ?w) ∧ IsA(?childworld, World, E, ?w) ∧ IsCounterFactualTo(?childworld, ?parentworld, E, ?w) ∧ IsA(?ag, Agent, E, ?w) ∧ Same(self, ?ag, ?childworld, E, ?w) ∧ IsA(?ev, Event, E, ?w) ∧ Occurs(?ev, ?parentworld, E, ?w) (.99)¿ Occurs(?ev, ?childworld, E, ?w)

For any world and for any focal event that happens in that world that involves an agent, the larger the causal chain of the focal event, the more likely the agent caused the focal event. This constraint captures the "zooming" phenomena described by N & K.

c_{10}: IsA(?parentworld, World, E, ?w) ∧ IsA(?childworld, World, E, ?w) ∧ IsCounterFactualTo(?childworld, ?parentworld, E, ?w) ∧ IsA(?ag, Agent, E, ?w) ∧ Same(self, ?ag, ?childworld, E, ?w) ∧ FocalEvent(?fe, ?childworld, E, ?w) ∧ IsA(?ce, Event, E, ?w) ∧ InCausalChain(?ce, ?fe, ?childworld, E, ?w) (.99)¿ Causes(?ag, ?fe, ?parentworld, E, ?w)

There are a few other constraints that capture mutual exclusivity relations between each event instance and agent instance as well. We now define the initial conditions of the vignette, which essentially is a description of what subjects read, plus some very basic background facts:

Events: IsA(glassFall, Event, E, R), IsA(armMotion, Event, E, R), IsA(twitch, Event, E, R), IsA(loudNoise, Event, E, R)
Agents: IsA(john, Agent, E, R)
Worlds: IsA(selfworld, World, E, R), IsA(otherworld, World, E, R)
For Mindreading: IsCounterFactualTo(otherworld, selfworld, E, R), Same(self, john, otherworld, E, R)
Percepts: Occurs(loudNoise, selfworld, E, R), FocalEvent(loudNoise, selfworld, E, R)

Once the loud noise is encountered as a focal event with John as a potential cause, simulation begins in order to explain the outcome. The loud noise occurs in the simulated world via the inheritance constraint c_9, and all of the antecedent events and causes are inferred by applying $c_6 - c_8$ to the simulated occurrence of the loud noise. Each of the antecedent events become focal events in the simulation, and causal chains for each are calculated via c_5 and c_6. Worlds having longer causal chains and not having agents as causes are penalized via c_{10}. Given the loud noise as the initially perceived event, the model produces the following output in the "best" (least penalized) world:

Best World: 260*, *cost: 2.0000004
Causes(armMotion, glassFall, selfworld, E, 260) ¿ true
Causes(glassFall, loudNoise, selfworld, E, 260) ¿ true
Causes(john, armMotion, selfworld, E, 260) ¿ true
Causes(john, glassFall, selfworld, E, 260) ¿ true
Causes(john, loudNoise, selfworld, E, 260) ¿ true
Causes(twitch, armMotion, selfworld, E, 260) ¿ true

On this particular parametrization of costs in the model, it seems as if John is blamed for everything from his arm motion all the way to the loud noise. By playing with the costs, we can adjust how many elements of the loud noise's causal chain John will be considered responsible for. As in N & K's experiments, we then give the model a set of inputs corresponding to the the question whether John was responsible for twitching. If we run the model with the same input as above except for the replacement of the percepts with Occurs(twitch, selfworld, E, R), FocalEvent(twitch, selfworld, E, R) we get the following output:

Best World: 213*, *cost: 0.2
Causes(twitch, twitch, selfworld, E, 213) ¿ false
Causes(john, twitch, selfworld, E, 213) ¿ false

In short, it seems as if we were able to capture the idea that more elaborate mental simulations make causal attribution to agents much more likely. These efforts are extremely preliminary, but nonetheless perhaps a sign that in subsequent work we shall be able to further model some of the more subtle influences on human attributions of responsibility.

5 Summary and Future Research

We've advanced the argument that cognitive science has much to offer machine ethicists and others seeking to build artificial moral agents. Ethicists might protest, by claiming that ethics has little to do with how people actually think, and that intuitions are irrelevant. After all, professional ethics has been and currently is an enterprise concerned deeply with the rational application of principles. To us, this seems to miss the wider point. Utilitarian, deontological, and virtue ethics resonate deeply with us (as humans) precisely because they often produce analyses that cohere with our intuitions. On the

other hand, they produce scores of paradoxes and conflicting recommendations when applied individually. We have suggested that the source of some of our intuitions about ethics-relevant concepts might reside in aspects of our cognitive architecture that weren't primarily designed to generate moral evaluations per se; but designed to generate predictions and explanations of agent behavior in mental-state terms via simulations. The upshot of our modeling work to date has been to reproduce the puzzling pattern of human judgments found by Nichols and Knobe regarding different conceptions of the "self." Our model suggests that perhaps we don't need multiple conceptions of the self to explain the data. Rather, our preliminary modeling exercise suggests that some of the patterns of judgments we see in the data can be explained via the interaction between judgments of responsibility and constraints on mindreading.

REFERENCES

[1] R. Arkin, *Governing Lethal Behavior in Autonomous Robots*, Chapman and Hall/CRC, 2009.

[2] P. Bello, 'Shared representations of belief and their effects on action selection: A preliminary computational cognitive model', in *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, eds., L. Carlson, C. Hoelscher, and T. Shipley, pp. 2997–3002. Cogntive Science Society, (2011).

[3] P. Bello, P. Bignoli, and N. Cassimatis, 'Attention and association explain the emergence of reasoning about false belief in young children', in *Proceedings of the 8th International Conference on Cognitive Modeling*, pp. 169–174, University of Michigan, Ann Arbor, MI, (2007).

[4] P. Bello and S. Bringsjord, 'Machine ethics, folks intuitions and the cognitive architecture of moral judgment', in *Proceedings of the AISB/IACAP World Congress 2012*, (submitted).

[5] P. Bello and M. Guarini, 'Introspection and mindreading as mental simulation', in *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, eds., S. Ohlsson and R. Ca, pp. 2022–2028, Austin TX, (2010). Cogntive Science Society.

[6] S. Bringsjord, K. Arkoudas, and P. Bello, 'Toward a general logicist methodology for engineering ethically correct robots', *IEEE Intelligent Systems*, **21**(4), 38–44, (2006).

[7] Nicholas L. Cassimatis, Arthi Murugesan, and Perrin G. Bignoli, 'Inference with relational theories over infinite domains.', in *FLAIRS Conference*, eds., H. Chad Lane and Hans W. Guesgen. AAAI Press, (2009).

[8] A. Goldman, *Simulating Minds:*, Oxford University Press, 2006.

[9] M. Guarini, 'Computational neural modeling and the philosophy of ethics', in *Machine Ethics*, eds., M. Anderson and S. Anderson, 316–334, Cambridge University Press, (2011).

[10] J. Knobe and S. Nichols, 'Free will and the bounds of the self', in *Oxford Handbook of Free Will: Second Edition*, ed., R. Kane, 530–554, Oxford University Press, (2011).

[11] B. McLaren, 'Computational models of ethical reasoning: Challenges, initial steps and future directions', in *Machine Ethics*, eds., M. Anderson and S. Anderson, 297–315, Cambridge University Press, (2011).

[12] Erik Mueller, *Commonsense Reasoning*, Morgan Kaufmann, 2006.

[13] S. Nichols and J. Knobe, 'Moral responsibility and determinism: the cognitive science of folk intuitions', in *Experimental Philosophy*, eds., J. Knobe and Nichols, 105–128, Oxford, (2008).

[14] A. Tversky and D. Kahneman, 'The framing of decisions and the psychology of choice.', *Science*, **211**, 453–458, (1981).