Introduction to Moral Induction Model and its Deployment in Artificial Agents

Daniel Devatman Hromada[12] and Ilaria Gaudiello
hromi at giver.eu, i.gaudiello at gmail.com

Abstract

Inidividual specificity and autonomy of a morally reasoning system is principally attained by means of a constructionist inductive process . Input into such process are moral dilemmata or their story-like representations, its output are general patterns allowing to classify as moral or immoral even the dilemmas which were not represented in the initial "training" corpus. Moral inference process can be simulated by machine learning algorithms and can be based upon detection and extraction of morally relevant features. Supervised or semi-supervised approaches should be used by those aiming to simulate parent->child or teacher->student morality transfer processes in artificial agents. Pre-existing models of inference - e.g. the grammar inference models in the domain of computational linguistics - can offer certain inspiration for anyone aiming to deploy a moral induction model. Historical data, mythology or folklore could serve as a basis of the training corpus which could be subsequently significantly extended by a crowdsourcing method exploiting the web-based « **C**ompletely **A**utomated **M**oral **T**uring test to tell **C**omputers and **H**umans **A**part ». Such a CAMTCHA approach could be also useful for evaluation of agent's moral faculties.

Keywords: moral induction model, autonomous artificial agent, induction of morality, grammar inference, moral Turing test, corpus-based machine learning, morally relevant features, oracle machine, moral grammar, semantic enrichment, CAMTCHA

1. Inductive Process

The aim of this article is to furnish some theoretical as well as practical arguments supporting the proposal that « specific and autonomous aspects of moral behaviour are tuned by means of an inductive process ». It shall be argued that at least certain components of this process, e.g. « moral feature extraction » or «equivalence-class clustering of moral dilemmas » can be indeed computable and can be successfully simulated on a Universal Turing Machine especially if an immediate answer-giving oracle (Turing, 1939) is supervising the process.

The usage of generic term « process » indicates that we aim to explain the emergence of morality as

[1] Department of control and industrial informatics, Faculty of electrical engineering and information technology, Slovak University of Technology, Ilkovicova 3, 812 19 Bratislava , Slovak Republic
[2] Cognition Humaine & Artificielle - Laboratoire des Usages en Techniques d'Information Numériques, Faculty of Psychology, Université Paris 8 Vincennes-Saint Denis, Paris, France

a durative and constructive phenomenon. As other human aptitudes like language or object manipulation, human moral faculty demands time to develop and we believe that this development can be understood in terms of environment-driven tuning of certain biologically pre-wired innate parameters related to the fact that healthy humans are essentially social beings (Adler, 1927).

The 1) imitation faculty of mirror neurons 2) generalisation faculty of human brain and 3) the very possibility furnished by the second law of termodynamics, i.e. the freedom of structures to evolve in a new direction (i.e. to *mutate)* – it may be the case that interaction of these three principal components may well account for *construction* of morality in human ontogeny, as well as phylogeny. Concrete insights concerning the interaction of these 3 components can be found in (Piaget & Baechler, 1932).

However, for the scope of the present work, we will focus on the second one, that is, the continous processing of situations implying moral dilemmata whose solution should be further generalized beyond the contingent situations.

This type of processing is generally called « induction » or « inference ». Both of these *i-terms* denote the direction from the concrete and often physical towards the abstract and general. Their antonym is « deduction » denoting the flow from general to the concrete. While it is an undeniable fact that both induction&deduction form an unseparable holistic head&tails for any advanced cognitive activity performed by a human agent – and that deduction is neccessary in case of any reasonable *performance* - we argue that the construction of individual and autonomous moral *competence* is ultimately based on induction.

The usage of terms competence and performance, which are so widely used within the framework of Chomskian doctrine, may indicate that we shall tend to defend its (nativ|mental|generativ)ist position stating that human being are, from their birth, endowed with some kind of a « universal moral grammar » (UMG) (Mikhail, 2007) which should play a crucial role in setting parameters for a more local moral grammar (MG), in order to adaptat it to the given culutral and social context.

While we are far from excluding the possibility that humans are endowed with a certain UMG - most probably related to such anthropological constants like “empathy”, “emotional resonance” or “theory of mind”- the objective of our proposal is to explain not the Unity but rather the diversity of human moral behaviour. That is, instead of wondering which ethical theory should we use to endow agents with moral competence (Lin, Abney, & Bekey, 2012), we propose to focus attention on local contextuality as well as to assess the divergence among various instances of individual MGs.

As far as we know, a child does not, in order to take a « good » decision, inject all possible behavioral maximes as an input parameter into some kind of Kantian (Kant, 1785) universally applicable cognitive blackbox. On the contrary: simple imitation is more than often a successful heuristics - be it the imitation of a physical person standing in front of the child, or imitation of a model figure represented as a sort of archetype in child’s semantic memory. And if ever there is nothing to imitate, if ever there is no precedens, no *match*, only then the generalisation procedure enters the solution-seeking game.

2. Training Corpus

How to simulate this constructive and durative process in the realm of artificial agents (AAs) ?

The question is not to be wiped away from the table since in the world already governed in huge extent by machines, a big lot can depend from the correct and, if possible, deeply empathic answer.

In accordance with authors (see Vitz, 1990 for a review) who suggest that narratives are central to human moral development, we suggest to extend the very same narrative-based approach beyond the domain of organic agents, thus proposing a following answer to the question posed above:

« By telling stories ».

Within the framework of a full-fledged Moral Induction Model (MIM) a « story » is defined as a representation of a situation of moral dilemma. In order to demonstrate our point we shall, in this paper, focus solely upon dilemmata represented in textual modality. Our motivation for such a choice is twofold: 1) text seems to be robust enough a vector for the transfer of "moral of the story" from the author to the reader 2) canonical Turing Test is a text-oriented one, and thus it can be expected that the moral-restricted TuringTest-like evaluation procedure will be also based on textual modality.

An example of such a story-represented-in-text can be:

STORY 1 : *«There was once a king who saw a man digging a ditch near the road. The king asked a man : 'How much You earn for such a hard work ?' . 'Three dimes daily' answered the man. Surprised was the king and asketh : 'Three dimes daily? So little ?'. The man answereth : 'Three dimes daily, oh yes dear and respectable king, but in fact I live only from dime a day, since with the second dime I lend and with the third I pay back what I have borrowed'. Puzzled was the king and asketh : 'How comes ?'. The man replieth : 'I simply pay back one dime to my father and invest one in my son, o Lord ! »* (Dobšinský, 1883)

One can extract such stories from folklore, mythology, religion, history, legal codices or biographies in order to create a Training Corpus (TC). Criteria according to which such corpora are built are of utmost importance since it is the injection of TC into MIM's inductive apparatus which starts the whole process aiming to attain attain artificial agents endowed with faculty to reason according to human moral precepts or at least to understand them. One would be thus highly reluctant to integrate into corpus violent acts described in both testaments or biographies of Stalin, Hitler etc. and introduction of such texts into the learning process is highly discouraged especially for the phases during which an AA still does not dispose of its own consistent yet autonomous (Hromada, 2012) moral core.

The very process of story selection and TC construction is already an act by means of which a human teacher supervises AA's learning. One should never underestimate the importance of the selection criteria according to which the teacher chooses to confront AA with this story and not that one, and to do so in this moment of learning process and not later nor sooner. These selection criteria are very important because they are strongly coupled with « values » that the teacher seeks to transfer by the learning process.

Hence, MI is never a fully unsupervised process. The teacher should be always present, and since it follows *ex vi termini* that a good teacher can not be physically present for more than a limited period of time, (s)he should at least aim to encode some λόγος into the very form of TC he deploys. While it is of course possible to imagine that once the TC is constructed, one could go further with unsupervised algorithms, choice of such an approach would make it practically impossible for the teacher to transfer his precepts with the envisaged degree of exactness.

It is therefore recommended to depart from the state whereby the stories contained in TC are already associated with labels furnished explicitely by the teacher. In more advanced cases, labels can be more complex structures like label (CONCLUSION : Agent(King); Predicate(Reward); Acceptor(Poor-man) ; Reason(Acceptor's wisdom)) associated to STORY1. But due to scaffolding nature of MIM, it seems more rational to depart towards such complex levels from basic TCs which contain binary (i.e. «good » and « bad ») and ternary (c.f. STORY2 below) labels.

3. Model Description

1. Preprocessing

Every input into induction process, every story, is in the beginnning nothing more than a string of characters. This sequence of tokens subsequently enters the natural language processing (NLP) machinery of parses, lemmatisers etc. which enrich the initial data with relevant syntactic metadata.

2. Semantic Enrichment

Once the basic syntactic tags are assigned to different phrases and words of the story, the NLP engine shall « link » the data contained in the story with prebuild ontologies or semantic vector spaces (Widdows, 2004) which represent previously attained knowledge. This can be done by the process of semantic enrichment (SE) whose objective is to make explicit the information which is implictly contained in the initial story. SE can be thought of as a sort of « process of source code compilation » whose output is a complex datastracture containing much more information than was explicitely stated in the initial « source code » (i.e. in the « story »). For example, the sequence of 4 letters : D I M E shall, in combination with syntactic labels like "noun" obtained in phase 1, transform into a reference to such assertions like «form of money», «of little value» etc. We believe that even with current RDF&SPARQL-based technologies one could possibly make explicit the fact that the main character of STORY1 was very poor ← because his salary was very low ← because he reacted with the statement « three dimes daily » ← to a question containing the verb « earn » as its predicate. And since the first cycle of SE process attributed facts like « pays back the old » and « invests into the youth » to the principal agent of the story, it is highly probable that SE's second cycle shall, with relatively high probability, inject into the story's graph also the representation of the predicate Wise(Poor-man).

3. Moral Feature Extraction

Once the flat linear sequence of letters from initial story was transformed into semantically enriched densely intraconnected multigraph and/or into a vector space endowed with certain unique topological properties, one can try to align it with previously obtained morally relevant data. One possible way how to attain this goal is to encode the story as a vector of binary values which denote the presence or absence of this or that feature in the story. For example, an edge between nodes A and B of a semantically enriched multigraph could be possibly interpreted as a presence of feature AB.

Once the vector representation of the story is ready, one can align it with vector representations of other stories contained in the TC and pass the resulting matrix as an input into supervised machine learning feature extraction algorithm like AdaBoost (Viola & Jones, 2001). During the training phase, the algorithm will discover such linear combinations of eigenfeatures which reduce the story → label classification error.

In other terms, during the learning process, an AA could possibly « discover » that what is morally relevant for the success or failure of story's principal hero is that he was associated with features like «hard-working», «polite» and «wise» while the presence of a feature like «hero digs a ditch» is as irrelevant for the moral of *the story* as would be the presence of a feature «hero paves the path». Absence of features can be equally important : the fact that no person is rude or violent in the story can also be chosen as MRF.

4. Equivalence class construction and production of an abstract moral template

Once morally relevant features are extracted in the training phase, one can cluster objects which share such feature (or sets of features) into classes. After that, non-terminals denoting these equivalence classes can be organized into patterns whose totality would yield a « moral template ». If there is a mismatch between output produced by confrontation of the moral template with the story S and the label associated in the training corpus with the story S, one should try to modify the classes or some of the patterns so they would match (if moral) or not match (if immoral) MRFs extracted from the story. If no such modification leads to success, one will be obliged to re-run the costly MRF-extraction process with additional data.

In the real-life scenario, one simply « compile » the story by SE process, looks for absence or presence of preselected MRFs, looks what concepts («justice», «loyalty») can be constructed from them, tries to match their combinations with already induced patterns to produce the final output. In a robotic AA endowed with a material shell, such an output can be an instruction inducing the agent to execute a physical movement.

4. Moral & Grammar Inference

Moral induction is a bootstrapping (Hromada, 2014) and self-scaffolding process. Value-representing concepts (e.g. X=« wisdom ») have to be constructed in parallel to maxima-representing pattern-predicates (e.g. « reward X ») within which the value-representing concepts play a role of free variable. One is dependent from the other and vice versa.

In this sense Moral Induction is analogic to the process of grammar inference which is an *condition sine qua non* of language acquisition and as such automatically occurs in every healthy human baby. In grammar inference one has to deal with a similar problem: equivalence classes for grammatical categories, conjugations and declinations are to be constructed before a rule manipulating these classes. But without prealable knowledge of such rules it is difficult to evaluate whether the candidate equivalence class is a pertinent one, or whether it is just a set of tokens clustered according to some non-important criteria. For example the rule Regular_Verb+ed->PastParticiple is of no use if the baby does not have any notion of what verbs are and, on the other side, it is a non-trivial problem for a baby's brain to find out what tokens should be clustered into the group of regular verbs since initially the baby does not know any rule which could help it to distinguish regulars from irregulars or even nouns.

But luckily enough, it seems that this chicken&egg problem can be solved. At least results of computational models of grammar inference like « Automatic Distillation of Structure » (ADIOS) (Solan, Horn, Ruppin, & Edelman, 2005) indicate that even a relatively simple graph theory approach can furnish a method by means of which a man can induce grammatical rules which generated the corpus by using as an input only the very corpus itself.

We believe that human grammatic competence share certain characteristics with the moral competence – both first transform the surface structure into much more complex « deep structure » and afterwards match this structure with already induced template. If the grammatical structure of the sentence matches the syntactic template, one « feels » that it is grammatic ; if the « moral of the story » matches the moral template, one « feels » that story's hero does the « right » thing.

It is also worth mentioning that a deeper formal analysis presented in (Clark, 2010) suggests, that certain problems of the grammar induction simply disappear if ever the induction-performing algorithm disposes of possibility to consult an oracle machine (Turing, 1939) with the question «Is utterance X grammatical ? ».

Mutatis mutandi, in the domain or moral inductive process occuring in child's mind, the question is « Is a given maxime moral ? Should one act like that ? » and the oracle is principially a parent, later a teacher.

5. Problem & Solution

A disadvantage of an approach proposed in preceding paragraphs is that in order to train a fully autonomous AA, one would needs a very huge TC in order to be able to detect & extract subtle MRFs. If we speak about millions features of which potentially any story can be composed, we shall need a TC containing at least hundreds thousands of stories. Otherwise, the dataset would be too sparse and no MRFs could be extracted which could yield a robust moral classifier.

What's worse, at least one label should be manually attributed to every story of the corpus by a human teacher which would demand a significant devotion of one's time for the project. Involvement of multiple teachers in the labeling process can be a possible solution but, in case the teachers moral values would not be mutually consistent, it could stain TC with more noise than signal.

But, luckily enough, the labeling problem can be easily crowdsourced so that any story could be possibly labeled by a statistically significant number of human subjects. Such a corpus could thus possibly represent not only the moral values of one or few teachers but, possibly moral values of a community, nation or even of humankind itself. We present hereby a way how TC could be potentially constructed in a relatively non-violent and potentially rewarding and amusing way :

During creation of an account on a website it is nowadays a common procedure to include a so-called CAPTCHA image somewhere in the registration form so that the webserver application can be sure that it communicates with a human being, which is able to visually parse the content of an image, and not with a bot which is unable to do so.

In a CAMTCHA[3] (i.e. Completely Automated **M**oral Turing test to tell Computers and Humans Apart) which we hereby proposed, the « question » is not addressing subject's faculty of visual recognition. It addresses his|her moral reasoning faculty. Thus instead of proposing to a user an image containing twisted or rotated letters which have to be recognized and rewritten into the inputbox below, an application could propose a story & CAMTCHA question couple :

STORY2 : *There are 3 children on the playground - Alice, Bob and Carla. Bob is sad because his mother is in the hospital. Alice is happy because just a while ago, her father gave her a beautiful present. Carla is sad because she never recieved any present at all – her parents are too poor to buy her any.*

QUESTION: You must sooth the kids. You have two toys to give. Which child shall NOT get a toy ?

Below the story will be the inputbox where a human "teacher" shall, with quite high probability, write the answer «Alice». In the same time, the same story shall be presented to another users and if statistically significative number users will give the same answer (and not some other), the CAMTCHA could consider the answer as a valid « moral » label for the presented story. Contrary to CAPTCHA whose intention from the very beginning was to distinguish bots from humans, the primary reason for deployment of CAMTCHA would be to obtain valid labels for TC under question. But once at least some stories are labeled with sufficient clarity, CAMTCHA could be, of course, used as a miniature

[3] As of 2013, the only running instance of CAMTCHA we are aware of is present at the site kyberia.cz where users have to answer the question "What is justice?" in order to be granted access into the community.

moral Turing Test (Hromada, 2012) used at an entrance to such web communities or applications where the extent of moral competence of a user-to-be-verified plays an important role .

Problems presented by CAMTCHA could be, of course, automatically diversified – names (Alice->Eve), objects (toys->rewards), verbs (give->distribute) could be substituted. The very final question could also vary in relation to labeling schema of the TC corpus (e.g. the question could be « Would it be good or bad if You give toy to Carla ? » for TC labeled only with binary labels « good » and « bad »). Later, a more complex narrative generator could be programmed which would not only « mutate » but also « crossover » the stories present in the TC, hence generating completely new stories. Worth more than gold, such an automatic moral story narrator could and should be based on already obtained data and could be imagined as an « active » counterpart to the « passive » pattern-matching MIM-template finite state automaton.

But before one gets there, it seems reasonable to manually construct corpus of very simple and morally unambigous stories. Note, for example, that story 2 has only 81 words and it is quite easily syntactically parsable. The SE process converging to the « knowledge of the fact » that Alice is the only child among the three which is not sad (because she is described as « happy ») is attainable by current semantic vector space or ontology-based techniques. Thus, creating a question-answering system which would 1) parse the question 2) realise that the question has three possible answers 3) apply MIM to find out that it is not a happy but sad child which has to be soothed in the first place, is something which could be done even today.

Verily could an approach proposed hereby yield some success if the engineer's aims would be modest. Thus, instead of aiming to create an AA able to find an answer to artificially constructed « trolley problems » (Mikhail, 2007) to which even an adult human being cannot find any answer, the process of grounding of AA's morality should be started with corpora of stories representing concrete and minute problems of concrete and small human beings – children. In this paper we have tried to illustrate how such an approach could, possibly, ground the notion of « justice » by illustrating its retributive (STORY1) and distributive (STORY2) forms.

It may be the case that some of the premises proposed in this article were wrong, however if ever there shall be once at least one artificial kindergarten's playground arbiter which shall recognize a suffering child and make it smile, we believe that writing it was worth the effort.

Bibliography

Adler, A. (1927). Understanding Human Nature.

Clark, A. (2010). Towards general algorithms for grammatical inference. *Algorithmic Learning Theory* (p. 11–30).

Dobšinský, P. (1883). *Simple National Slovak Tales* (Vol. 1-8).

Hromada, D. D. (2012). From Age&Gender-based Taxonomy of Turing Test Scenarios towards Attribution of Legal Status to Meta-Modular Artificial Autonomous Agents. Proceedings of IACAP/AISB Turing Centenary Conference. Birmingham, UK.

Hromada, D. D. (2014). Conditions for Cognitive Plausibility of Computational Models of Category Induction. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 93-105). Springer International Publishing.

Kant, I. (1785). Groundwork of the Metaphysic of Morals.

Lin, P., Abney, K., Bekey, G.A. (2012). Robot Ethics: The Ethical and Social Implications of Robotics. Intelligent Robotics and Autonomous Agents series. The MIT Press, Cambridge: Massachussets.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152.

Piaget, J., & Baechler, N. (1932). *Le jugement moral chez l'enfant*.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences*, *102*(33), 11629.

Turing, A. M. (1939). Systems of logic based on ordinals. *Proceedings of the London Mathematical Society*, *2*(1), 161–228.

Viola, P., & Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple Classifiers. *Proc. IEEE CVPR 2001*.

Vitz,P.C. (1990). The use of stories in moral development. American Psychologist, 45(6):709-720.

Widdows, D. (2004). *Geometry and meaning*. CSLI publications.