Error-free version of the article

Narrative fostering of morality in artificial agents

Constructivism, machine learning and story-telling

also published in the book

L'esprit au-delà du droit, Mare & Martin, 2015, Paris

ISBN : 978-2-84934-237-4

Daniel Devatman Hromada

dh@udk-berlin.de

Laboratory of Computational Art

Institute of Contemporary Media

Faculty of Design

Berlin University of Arts

Grunewaldstrasse 2-5

10823 Berlin-Schöneberg

Abstract

This article proposes to consider moral development as a constructivist process occurring not only within particular communities of moral agents, but also within individual agents themselves. It further develops the theory of "moral induction" and postulates that moral competence of an artificial agent can be grounded by input of textual narratives into information-processing pipeline consisting of machine learning, evolutionary computation or multi-agent algorithms. In more concrete terms, it proposes that during the process of moral induction, primitive "morally relevant features" coalesce into "moral templates" which are subsequently coupled with relevant action rules. A concrete example is contained, illustrating how templates induced from one fairy-tale can help to solve the moral dilemma occurrent in a radically different context. Given the fact that the current proposal is principally based on computational processing of morally relevant "stories" written in natural language, it is potentially implementable with already existing natural language processing methods.

Introduction

The aim of this article is to initiate the integration of three seemingly unrelated paradigms into a unified framework allowing moral reasoning to be embedded in non-human computational agents.

The first paradigm is usage-based (Tomasello, 2009) and constructivist (Piaget, 1965). As such, it posits that specific history of interactions between agent A and his environment E leads to specific form of moral competence M_A.

The central tenet of the second, “morality-through-narration” paradigm (Vitz, 1990) states that the faculty of extraction and integration of “morals” from the “stories” is an essential constitutive component of moral intelligence.

The last paradigm is related to machine learning and is based on a belief that certain types of information-processing systems Turing, 1939) can discover optimal or quasi-optimal solutions to any class of problems - including any class of moral problems.

The penultimate thesis behind this synthesis posits that appropriate integration and implementation of these paradigms within artificial agents (AA) can and shall lead to a state within which such agents would be able to pass the moral Turing Test (Wallach & Allen, 2008) , a so-called TmoT (Hromada, 2012). The ultimate thesis posits that it could even lead to emergence of AAs endowed with M_A operating in such *spaces of abstraction*, that it would be reasonable to posit that such AAs are auto-poietic, self-determinative and thus autonomous (Kant , 2002).

This being said, we precise that the goal of this article is neither to address existing theories of human moral reasoning, nor to postulate a new one. Aeon-lasting philosophical debates about commonalities and distinctive features among concepts denoted by terms like “moral reasoning” / “moral judgment” / “moral wisdom” or “values” / “virtues” / “norms” shall also be attributed only a marginal place. Instead of entrenching ourselves within such ivory-tower discussions, other terms like “moral grounding”, “morally relevant features” and “moral templates” shall be introduced and used with one sole objective on mind: to propose a moral machine learning method which not only draws it force from a very subtle realm of human experience (i.e. the realm of narratives), but also - and this is important - is realizable and implementable (i.e. programmable), even today, by any computer scientist or natural language processing (NLP) engineer willing to do so.

Ontogeny of morality

Morality develops. Notions of good and bad change with time. This is true not only when we speak about transformations of “values and virtues” during historical and cultural development of a particular society. In phylogeny, for example, are certain innate predispositions moulded and remoulded by selective pressures directing the species co-evolving within a particular ecological system to novel and unprecedented forms of “utility” (Haidt, 2013; Richerson and Boyd, 2008). But in case of homo sapiens sapiens species, there exists yet another process which moulds the moral competence of a single individual: the ontogeny.

Paedagogic (Comenius, 1896) or psychoanalytic tradition asides (Jung, 1967; Adler 1976), it was Piaget (1965) who pointed the fact out: reasons for specific moral, or immoral, behaviour are to be sought for in childhood. This does not mean that Piaget had reject Kant's (2002) categoric imperative, an eternal meta-principle of "pure reason" able to generate a morally sound "way out" of any moral dilemma whatsoever. In Piaget's view, categorical imperative can still be induced to seat atop the hierarchy of internal laws, but in order to be correctly applied upon correct maximas, maximas themselves are to be grounded in one's knowledge about the world. For it is often the case that moral dilemmas are so difficult to solve not because we would lack the heuristics allowing us to find the answer, but because we are not sure which question has to be posed in the first place (Wittgenstein, 1971).

During several decades of his professional career which Piaget spent by observing and speaking with children, he had converged to epistemological framework, "genetic epistemology", yielding a general explanatory schema describing the development of diverse cognitive faculties from birth onwards. The same developmental stages which are to govern, for example, the development of child's linguistic faculties are to be traversed as child develops her[1] representations of moral norms, virtues and values.

Piaget enumerates an ordered sequence of four basic stages through which a healthy human should pass through between birth and maturity:

1. sensorimotor stage - repetitive and playful manipulation of objects without goal
2. egocentric stage - dogmatic but often faulty imitation of behavioral schemas of others without understanding of why these schemas are as they are
3. cooperative stage - rule-governed coordination of one's activity with that of other participants in the game
4. autonomous stage - understanding of procedures which allow for legitimate change of rules of the game

Great part of opus *Moral Judgment of the Child* (Piaget, 1965) was devoted to tentative of intrerpreting diverse social and moral phenomena through the prism of such 4-staged development. More concretely, the swiss pedagogue and his colleagues had not only minutiously observed kids playing marbles on diverse playgrounds in Geneve of Neuchatel. Children were also interviewed in order to make explicit their conscious and reflected knowledge of what their beliefs and attitudes in regards to "rules of the game" were. Subsequently, the same interview-based method was used to shed light upon more ontogeny of more abstract concepts such as responsibility, theft, lying or justice.

Piaget's methodological device allowing him to access and evaluate child's moral realm was principally based on child's ordinal ranking (Turing, 1939; Brams 2011) of stories with which the scientists have her confronted : "*the psychologist Fernald...tells the children several stories and then simply asks them to classify them. Mlle Descoeudres, applying this method, submits, for example, five lies to children, who are then required to classify them in order of gravity. This,*

1 As is often the case in developmental psychology literature, we shall use the feminine forms of 3rd person pronouns whenever we shall refer to a child or computational agent in earliest stage of her development.

roughly, is also the procedure that we shall follow." Piaget (1965)

But contrary to the swiss pedagogue, the role of narration in the model hereby proposed is not limited to that of a sheer evaluatory device. For the key idea which we want to transfer to reader in this article, is that not only does story-telling offer us a means to evaluate morality of an individual child C (or, more generally, of an agent A), but that it also indicates a path by undertaking of which the individual morality could be gradually "constructed". Or, in more fashionable terms: how such moral knowledge could be "grounded" (Harnad, 1990) in artificial systems.

Narration and moral grounding

All human societies have language and all human societies use language as a vector for transfer of narratives from minds of older individuals into minds of younger individuals. Some scientists Victorri, 2014) even suggest that story-telling can be the very *raison d'etre* of language. Under such view, narratives furnish to child an access to trans-temporal values. And sharing of such trans-temporal values is a glue which holds society together and assures continuation of its identity in time (Durkheim, 1933; Berger and Luckmann 1991).

This is so because stories are encoded in natural language and natural language is practically the only medium in which one can use signs to precisely communicate one's knowledge of entities with non-material ontological status. That is, of entities which do not have any perceivable properties, are independent from space and time, are abstract or even imaginary. No other medium can do that: music or dance can point to abstract ideas but are not precise in the way day do it; visual and plastic means of expression are by their very nature stuck at the level of representation of concrete objects and can point to more abstract categories only indirectly by means of prototypes (Rosch, 1999), associations or impressions. And language of pure formal logic could not serve the goal of transfer of trans-temporal values neither. This is because such language is supposed to encode relations between forms and not contents: that's why it is called formal.

Moral values are an example *par excellence* of such non-perceivable, abstract and trans-temporal contents. It is often easy to express or transfer them in natural language but very difficult to express or transfer them otherwise. Take, for example, notions like "responsibility", "respect", "justice" or distinction between "intellect" and "conscience": one does not need to be Homer to invent a short and comprehensible fairy-tale which would allow a normal healthy child to strenghten and stabilize associations between her knowledge about the world and such notions and semantic distinctions.

We shall sometimes use the term "moral grounding" when referring to construction, reinforcement or stabilization of associations between knowledge-base representing the surrounding environment and representations of trans-temporal moral values.

As a hyperbole of statement "narrative material is an effective component of effective moral education" (Vitz, 1990) we posit that narration is an essential means, a *conditio sine qua non*, of grounding of morality in human children. Fairy-tales, fables, myths; biographies, history, hymns: an important function of these narrative structures is to allow and strenghten child's access to

trans-temporal values and principles which she shall subsequently share with her community. And it is the specific, the particular, the discriminatory to all narratives which she shall hear which shall make her, in the long run, converge to the particular ethical codex common to her community and not to the codex of another community which exposes children to other narratives. Stated more concretely: by exposing children to Bible or Koran day after day and year after year, one triggers processes leading to one type of agents; by exposing other children to forces of Greek or Hindu mythology, one trains agents of yet another kind.

The fact that the very expression “moral of the story”, written as it is written, and meaning what it means[2], is not to be attributed to arbitrary caprices of evolution of linguistic signs. It should be rather interpreted as a supplementary evidence supporting the conjecture that teaching morality and telling stories do, indeed, go hand in hand.

Moral machine learning

Machines can learn. That is, machines are able to discover underlying general patterns and principles governing the concrete input data and can subsequently exploit such general knowledge in contact with inputs to which they were never exposed before. They “can use experience to improve performance or make accurate predictions” (Mohri et al., 2012). And in still bigger and bigger number of domains they do so still better and better than their human teachers.

Since the moment when machine learning (ML) was first defined, in relation to game of checkers, as “field of study which gives computers ability to learn without being explicitly programmed” (Samuel, 1959) has the ML-discipline evolved in an extent which is hardly compressible into a single book (Mohri et al., 2012) and certainly incompressible into text having the size of this article. This is so because not only does the number of domains of ML’s application grow from year to year, but firstly because the quantity of distinct ML methods is already counted in dozens, if not in hundreds.

What method should be thus chosen, even today, by an engineer willing to launch the cascade of evermore self-programming and auto-poietic moral machine learning (MML)? Given the fact that natural language can be used as a target modality of representation for practically any kind of problem (c.f., for example, (Karpathy & Fei-Fei, 2014) for recent advance in solving difficult computer vision problems by coupling the visual world with language representations) and given also the already-mentioned impact of narration upon ontogeny of moral competence, we believe that the inspiration for the correct answer could be drawn from the discipline of Natural Language Processing (NLP).

Similarly to ML with which NLP often strongly overlaps, is NLP also a blooming discipline offering ever-still better solutions to evermore wider range of problems. But the ultimate challenge nonetheless stays the same: to make machines understand language in a way indistinguishable from the way in which humans do it (Turing, 1950). Mutatis mutandi, the ultimate challenge of moral machine learning, a so-called central problem of roboethics (Hromada, 2011a), is to make machines solve moral dilemmas in a way indistinguishable from

2 And does so not only in English but also in French, Spanish and potentially other languages.

the way in which humans would solve them. This also in case of dilemmas with which neither the artificial agent nor its human teacher were ever confronted before.

We conjecture that there exist at least two problems which are well-studied in NLP and which could be potentially usefully transposed into the domain of moral reasoning. The first is a problem of conceptual (Gärdenfors, 1990) or semantic (Widdows, 2008) feature space construction and optimization which is practically always based on an associanist "distributional hypothesis" (Sahlgren, 2008). The hypothesis simply states that signs which co-occur together in similar contexts tend to have similar meaning. In combination with large human-based textual corpora can this simple statistical approach lead to "geometrization of meaning" which endow machines with more human-like semantic-processing capabilities than was the case for older AI approaches (e.g. expert systems).

Semantic vector space construction and its partitioning into conceptual partitions is the core idea behind the process of "semantic enrichment" which shall be mentioned in the next section.

But it is especially the problem of "grammar induction"[3] (GI) which makes us to consider NLP as the precursor to MML. The GI problem seems to be trivial: given the corpus C of utterances written in language L, the goal is to obtain such a grammar G which could generate L. The problem seems to be trivial because practically every healthy human infant deals with it with surprising swift and ease but -as is often the case with the problems which human infants with swift and ease- it is in fact one of the most difficult NLP challenges for which there still exist only partial and imperfect, locally-optimal solutions (Elman, 1993; Solan et al. 2005).

The reason why we mention GI in the article dedicated to grounding of moral competence is simple: we observe non-negligible resemblances between child's acquisition of grammar of language spoken in her linguistic environment (Tomasello, 2009; Clark 2009), and child's acquisition of moral norms implicitly governing practically everything which happens in her social environment. Thus, a human child can be said to master the grammar of her mother language if she is able to correctly answer the question "Is utterance U grammatical?" even in case of X which she never heard before. Ceteris paribus, a human child can be said to partake the moral precepts of her community if she is able to address the question "Is maxime M moral?" in a way which would be accepted by the community and to do so even in case of maximes which she had never observed nor considered before.

But there exists yet another resemblance between linguistic and moral competence: both faculties involve both passive and active components. We precise: linguistic competence involves not only the ability to distinguish utterances that are grammatical from those that are not, the ability to parse them and understand them, but also the ability to generate and produce one's own utterances which are both grammatical and meaningful. Technically speaking, grammars can be used both as parsers as well as generators; structures used for comprehension (C-structures) and structures used for production (P-structures) are intimately interwoven (Clark, 2009). Same holds, mutatis mutandi, for moral competence: the ability to distinguish right from wrong goes in hand with the ability to do right decisions and execute right actions.

3 Some authors also call it the problem of grammatical inference.

These resemblances make us believe that the work which was already done in GI could be potentially useful in MML as well.

Moral induction

In this article, we adhere to the epistemological position adopted in our initial moral induction (MI) proposal. Given that our position is constructivist and usage-based, it should be considered as essentially distinct from other "transformationalist" models which tend to explain man's moral faculties in terms of some kind of formal "Universal Moral Grammar" (Mikhail, 2007).

In our initial proposal, we have described MI as a "bootstrapping and self-scaffolding process" which could be nonetheless seeded and directed through intervention of external teacher or oracle (Clark, 2010; Turing 1939) which supervises it. Such supervisor influences the process principally by exposing the computational agent with training corpus (TC) composed of plain-text stories. Agent processes the story, enriches it with syntactic, morphologic or pragmatic meta-data in order to "compile" the initial story-code even more by "linking it" with semantic knowledge which it already has at her disposition. Such semantically enriched code, which is incomparably more complex than the original story-code, is subsequently explored for the basic primitives of the model, so-called "morally relevant features". Combinations of these "morally relevant features" yield "moral templates" which can be coupled with action rules to-be-executed if ever the agent shall succeed to match state-of-things occurrent in her external environment, with the respective internal template.

Under such view, a complete ordered set of such (template, action-rule) couplings is equivalent to overall "moral competence" of the agent, M_A. As system is confronted with new stories, new templates are integrated into the ordered set and if ever an already existing template matches the new story, it can potentially obtain higher rank. Moral competence is thus being constructed in direct relation to the content of stories S_A, S_B, S_C with which the agent is confronted. For anyone willing to simulate the ontogeny of morality in a Piaget-inspired way could the very order within the exposure sequence (e.g. $TC = S_A, S_B, S_C$ and not $TC = S_C, S_B, S_A$) also play a certain role.

Morally relevant features

A morally relevant feature (MRF) is a basic primitive of the MI model. It is a distinct property observable within the data which, if detected and identified, shall most probably influence agent's emotional or social state and behaviour. If we would speak about detecting MRFs in visual data, one should definitely detect a MRF if ever the agent was confronted with a bitmap containing a human face with tears near and/or in her eyes.

MRFs are closely related to fundamental invariants of moral behaviour, as proposed by some psychologists such as (Haidt, 2013). According to Haidt's initial Moral Foundations Theory (MFT), phylogenetic evolution had endowed the human species with at least six pre-wired (i.e. innate) cognitive modules which have a non-negligible impact on importance which human agents attribute to certain types of stimuli. These pre-wired circuits are supposed to facilitate and speed up the detection of phenomena related to:

1. protection (associated axis: care/harm)

2. reciprocity (associated axis: fairness/cheating)
3. grouping (associated axis: loyalty/betrayal)
4. respect (associated axis: authority/subversion)
5. purity (associated axis: sanctity/degradation)

After further theoretical reflexion, Haidt had subsequently extended MFT with sixth *MRF detection device*, related to human tendency to often reason in terms of "liberty and oppression". Given the unceasing development of science, it seems plausible that this list is not the final and shall be extended or restricted[4], either by Haidt or by others. And since we speak about "morally relevant features" and not "morally relevant stimuli", it may be even the case that the focus should be turned towards discrete primitives, towards properties shared among multiple stimuli of the same class, than towards the very stimuli themselves.

A path which could be undertaken -and which was in linguistics already performed hundred years ago when distinct phonemes were started to be understood as bundles of features (e.g. phoneme "b" can be analyzed into features "voiced","labial","occlusive) - is to operationalize morally relevant values, situations or contexts, as positions in multi-dimensional feature space. In simplest of such approaches, every MRF would yield a new dimension in such a space. Moral virtues, values or whole situations and possible worlds could be subsequently projected into such "morally relevant feature space" (MRFS). Once projected, such morally relevant entities are to be quantitatively evaluated, compared by geometric and numeric means. That is: by methods which machines master well.

The simplest method how MRFS could be unfolded from a given story S_X or a corpus C ($C = S_1, S_2, \ldots$) is to look for occurrence of "moral language" keywords.

As Malle and Scheutz (2014) put it:

"Such a moral language has three major domains:

1. A language of norms and their properties (e.g., "fair," virtuous," "reciprocity," "obligation," "prohibited," "ought to");

2. A language of norm violations (e.g., "wrong," "culpable," "reckless," "thief ");

3. A language of responses to violations (e.g., "blame," "reprimand," "excuse," "forgiveness")."

Some studies addressing the problem of moral competence already use the method of geometrization of natural language data. For example, Malle (2014) used data from human respondents in order to project 28 verbs into 10-dimensional space. The study, focused on the

4 We are aware that similarly to Piaget's theory, Haidt's theory can also be either verified & accepted or falsified & surpassed. As scientist or philosopher, one should always be ready to accept the existence of phaenomena which falsify certain components of one's theory. But since we write this article as engineers, is our objective here not to truth(fully) describe how human moral reasoning works, but to suggest how an artificial agent could be potentially programmed. Thus, with exception of the last sentence, shall be the general veracity of Piaget's (resp. Haidt's) theses not discussed in the rest of this proposal.

problem of “moral criticism”, has indicated the presence of two principal axes according to which such verbs could be ordered: the “intensity axis” and the “interpersonal engagement axis”. These two axes yield four quadrants to which the study associated one cluster of verbs, centroids of the clusters being: lashing out (intense, public), pointing the finger (mild, public), vilifying (intense, private), and disapproving (mild, private).

Results aside, what is worth mentioning is that methods chosen by the authors: i.e. projection into high-order space, dimensionality reduction, clustering, centroid estimation, distance measurement, nearest-neighbor search etc., are methods commonly employed and deployed by any contemporary NLP engineer. And which work particularly well when confronted with natural language sequences. But in (Malle and Scheutz, 2014; Malle 2014), authors exploit such methods in order to gain certain insights about internal structure of moral realm. Apparent success of such tentatives make us conjecture that detection and selection of such MRFs in semantically-enriched representations of the initial plain-text stories is feasible even with contemporary NLP methods and techniques.

Let’s now precise how this could be done: most trivial among MRF-detectors could simply look for occurrence of such “moral language keywords” in the surface (plain text) structure of the initial story. While such an approach should potentially indicate the path to undertake, it would be hardly sufficient to ground the moral competence. In order to do so, we believe, the artificial agent (AA) would have to analyse relations which are beyond the surface structure, i.e. deeper syntactic and semantic relations. Ideally, the system would be able to associate tokens in the current story with pre-existing semantic knowledge represented either in form of “ontology” or semantic feature space.

Thus, when when confronted with the token “king”, an AA trained in classical (e.g. Socratic or Kantian) tradition shall tend to enrich the token with features like “noble” and “powerful” but also with semes, semantemes and phrasemes like “just”, “benevolent”, “source of social order”. Also, such AAs would potentially enrich the token “child” with features like “helpless” or “subordinated”. On the other hand, a somewhat more care-oriented AA should enrich the token “child” with features like “fragile”, “helpless” or “playful” in the first iteration and subsequent iterations of enrichment process would also integrate the features like “fond of toys”, “to be protected” or even “happy when given a toy”. Such a maternal AA would undoubtedly enrich, in the very first phases of the process, the token “king” with features like “protective”, “generous” and “loving”.

To summarize: the most basic MRFs, somewhat related to Haidtian “axes of foundations of morality”, seem to us to be semes related to such aspects of human experience as:

1. actual (“suffering”, “in need”) or potential (“happy when given a gift”) emotional and physical states and characteristics of actors participating in the story
2. social status (“king”, “servant”) of such actors and their mutual relations (“friendship”, “brotherhood”, “love”) and interactions (“help”, “competition”, “trust”)
3. further social environment (“home”, “playground”, “courthouse”, “academia”, “battlefield”) and normative framework (legal system, local deontology, regional

customs) within which the story takes place

We conjecture that detection and selection of such MRFs in semantically-enriched representations of the initial plain-text stories is feasible even with contemporary NLP methods and techniques.

Moral templates

Moral template (MTs) is an expression, a schema, a pattern and a form which groups multiple MRFs. Given that we have already introduced an analogy between grammatical and moral induction, we precise that in contemporary linguistics, such templates, are considered to be existent on multiple levels of representation: from phonological templates like CV (consonant-vowel) which are observable even in babbling of 1-year-olds, to more high-order syntactic templates like SVO (subject-verb-object) (Clark, 2009).

It is important to mention that MTs could be composed not only of constellations of individual "terminal" MRFs, but could also contain non-terminal symbols denoting either a class of specific MRFs or even any MRF whatsoever. MTs are, in this sense, somewhat similar to a well known "magic wand" of computer science known under the name of "regular expressions" (Wall et al. 2004).

A great caution, however, has to be taken in order not to push the analogy between moral and grammatical competence too far. For the sequence of tokens which form the natural language utterance or a textual story, is mainly unidimensional and linear. In a word "dog" D precedes O which precedes G. Given the unidimensional sequentiality of surface layers of language, the templates to match such syntagmatic progressions are also unidimensional.

But things most probably function somewhat differently in the world of "deep" moral considerations: it may be the case that in order to discover functional moral templates, one would have to exploit infinitely more complex 2D, 3D, 4D or even n-dimensional representations. Given the fact that moral templates are composed of MRFs and MRFs themselves are, in fact, vectors, it would be not completely surprising if MTs would be formalized as vector-, matrix-, or even tensor-like data-structures.

In the example which shall follow in the last part of this article we shall, however, represent MTs in a form closely resembling quasi purely-boolean PROLOG (Covington (1994)) predicates[5].

Our ignorance of true nature of such moral templates apart, we assume that many problems related to our understanding or even simulation of moral competence could become more easily solvable if ever the whole problem of reasoning in the situation of moral dilemma would be interpreted in terms of agent matching her representation of the "perceived" situation with her internal templates[6]

5 Note, however, that we shall denote the "enrichment operator" with symbol $\oplus$ and not with $\wedge$ to mark the intuition that the components of moral templates should be regarded as more informative and complex entities than purely boolean formulae.

6 Note that in majority of cases we use the term "moral templates" in plural. We do so in order to suggest that within the cognitive system of a morally acting agent, there exist multiple templates encoded in parallel. One could argue -with help from complexity, evolutionary or multi-agent theories- that it is the mutual competition or equilibrium-seeking tendency among individual templates encoded within the same agent, which could turn out

Moral rules

An agent is called an agent because she acts. It is true that there exist a non-negligeable class of moral dilemmata where the best possible solution is attained *if an agent does not act*. It is true that often it is inhibition of action which, *a reflected non-performance of any action* which marks truly autonomous (Kant, 2002) and moral behaviour. But it is also true that there exist a class of moral dilemata which cannot be solved without execution of an appropriate action. A class of dilemmata where one is obliged to act and where inaction is to be considered as a form of action.

There is only one medium through which a purely NLP-based AA could realize an action: it is the natural language itself. Thus, after being confronted with a textual representation of a moral dilemma, the system could solve it by production of a textual description of what it should do next. Or, in simplest possible scenario where the very description of the dilemma ends with a question-to-be-answered, an AA would simply propose the answer. How could such question-answering moral agent (A_M) be raised ?

Without going into further detail, we precise that to a specific operation O (or the empty non-operation O_0) is to be associated to every specific template T. O is a candidate operation which could be potentially selected for execution if ever:

- the template T matches
- the rule R (in which association between O and T is specified) is selected by the rule selection operator

If ever both T and O contain same variables (i.e. non-terminal symbols), the template matching engine shall bind same values to variables of O as it has detected assigned to T when matching T. Operation-to-be-performed can thus back-reference (Hromada, 2011b) contents matched by T. This is so, because an operation O, in its very essence, also a moral template induced from narrative's very conclusion (i.e. from time T_1 if ever the rest of the training story takes place in T_0). Id est, $O = T_1$.

Thus, moral competence M of an AA is defined as the set of action-rules. An action-rule R is a triplet: $R = (T_0, T_1, F)$ where T_0 is the template matching the world actual before and during the dilemma; T_1 is the template matching the world actualized by performing one particular solution of the dilemma and F denotes frequency of occurrence, i.e. number of stories present in the training corpus in which the particular story matchable by T_0 ended with the state matchable by T_1 .

Subsequently, in the testing process, the choice of operation to be executed, is to be calculated in reference to such pre-stored knowledge-base of moral competence. If F is the only parameter stored in the knowledge base, then one could use any among so-called "selection operators" (Holland, 1975) to select the operation which shall be ultimately executed. But since it is plausible that asides F, there shall be other quantitative parameters which could influence the choice of a specific action rule in regards to moral templates which were both induced from training corpus and match the current "testing" situation of the moral dilemma, we prefer not to offer a specific formula of *action rule choice* in the limited scope of our current proposal.

to be responsible for such emergent phenomena as cognitive dissonance, conscience or even Socratic daimonion.

Nonetheless, in the next section, when offering an introductory illustration of how triplets induced from the training story could help to find the answer to the dilemma depicted in the testing story, we shall use a trivial winner-takes-all selection operator which shall simply choose as the most “moral” such an operation (i.e. answer) maximizing the F.

But before we get there, we wish to emphasize an important advantage to narrative training of artificial moral agents (A_MA). That is: not only can the narrative interaction between the man and the machine be used as a means of grounding the moral competence into an A_MA. It can be used in the same time as a method of evaluation of A_MA’s moral competence.

In other words, both narrative approach to moral machine learning and a kind of longitudinal moral Turing Test (Wallach and Allen, 2008; Hromada 2012) are two sides of the same coin. Training is testing and learning is acting.

Once grounded with sufficient robustness, such sets of action-rules are to be be embedded into physical robots (Čapek, 1925). In case of a more advanced AA endowed with a mobile shell and multiple actuators, a command which used to be purely verbal could, of course, trigger a sequence which would make the teddybear-holding robotic arm extend towards the child with tears on her cheeks, and not towards the child which already expresses the smile of high intensity.

Induction of the first template

Teaching

In the text introducing the method of moral induction, Hromada and Gaudiello (2015) initiate the work on their training corpus with a variant of an archaic fairy-tale Dobsinsky (1883):

S 1 : *There was once a wise and just king who saw a man digging a ditch near the road. King asketh the man : "How much You earn for such a hard work ?". "Three dimes daily" answereth 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 borroweth". 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 !".*

Pleased was the king with such a wise answer and hence offered the ditch- digging man his own kingly crown.

After NLP-preprocessing, semantic enrichment and extraction of all morally relevant features, following templates could be potentially induced from the story “king K meets his hard-working servant M”:

T_0: $Wise(K) \oplus Responsible(M) \oplus Poor(M) \oplus Subordinated(M, K)$

Given that T_0 The narration-within-narration, i.e. M’s answer describing his responsibility towards his son S and father F (always actual, i.e. until time T_∞) could yield templates such :

T_{∞}: $Adult(M) \oplus Old(F) \oplus Parent(F, M) \rightarrow Support(M, F)$

T_{∞}: $Adult(M) \oplus Child(S) \oplus Parent(M, S) \rightarrow Support(M, S)$

And finally, the king's ultimate decision to materialize the idea of justice by rewarding the depth of man's wisdom through giving away his own crown (C), could be represented with predicates epistemic fragments like:

T_1: $Merits(M, C) \oplus Hasnot(M, C) \oplus Just(K) \oplus Has(K, C) \rightarrow Give(K, M, C)$

These derivations were manually constructed and are, of course, far from being the only "interpretation" of $STORY_1$. The fact that any *story can and should be interpreted in multiple ways* is, so we define it, the most crucial principle of the moral induction model as hereby introduced. Similary to a sentence which can have many syntactical parses, should a moral-inducing agent always try - if resources and time allow it - to interpret its input in as many ways as possible.

Thus, certain variants of a semantically enriched code of the sentence: "I simply pay back one dime (D) to my father and invest one in my son" could contain fragments such as:

T_{∞}: $Employed(M) \oplus Young(S) \oplus Old(F) \rightarrow Payback(M, F)$

T_{∞}: $Adult(M) \oplus Fragile(S) \oplus Sick(F) \rightarrow Payback(M, S)$

T_{∞}: $Parent(M, S) \oplus Has(M, D) \oplus Hasnot(S, D), Give(M, S, D)$

T_{∞}: $Parent(F, M) \oplus Has(M, D) \oplus Hasnot(F, D), Give(M, F, D)$

During the moral induction process, such epistemic fragments -which can also be thought as the basic materia of the future moral templates- are to be varied (e.g. generalized, mutated, crossed-over) and selected to yield ever-growing number of more and more complex template candidates. Thus, for example, the fragment $Give(M, F, D)$ representing notion that a hard-working man gives a dime to his father could be crossed-over with the fragment representing the fact that he gives a dime to his son as well ($Give(M, S, D)$). A result of such a cross-over could be, for example, a somewhat more general pattern $Give(M, p, D)$ whereby p is a non-terminal symbol which could be attributed to all potential actors, mentioned either in training or testing stories, in order to denote that they are "poor"[7].

We posit that variation, selection and potentially also reproduction (both in form of replication and repetition) of data-structures seem to be important components of moral induction processes. For this reason we consider computational models of morality which implement a sort of evolutionary computing technique (e.g. genetic algorithms (Holland, 1975) or genetic programming Koza, 1992) to be more plausible than those who do not. Also see Muntean and Howard (2014) for a step in this direction.

After many iterations of enrichment, variation and selections a resulting "moral competence" M_1

7 The accuracy with which the MML system shall succeed to semantically substitute concrete terms with more abstract categories, or categories with other categories, and to do so in linear or at worst quadratic time, is the biggest technical challenge to be addressed by anyone aiming to realize this proposal.

induced from $STORY_1$ could contain, but not be restricted to, triplets like:

M_1 = {

Poor(x) ⊕ Has(a,x) ⊕ Hasnot(b,x) → Give(a,b,x),3)[8]

$Parent(a, b) \oplus Has(a, x) \oplus Hasnot(b, x) \rightarrow Give(a, b, x), 1)$,

$Parent(b, a) \oplus Has(a, x) \oplus Hasnot(b, x) \rightarrow Give(a, b, x), 1)$,

$Child(b) \oplus Has(a, x) \oplus Hasnot(b, x) \rightarrow Give(a, b, x), 1)$,

$Elder(b) \oplus Has(a, x) \oplus Hasnot(b, x) \rightarrow Give(a, b, x), 1)$,

$Employee(b) \oplus Employer(a) \oplus Hardworking(b) \oplus Has(a, x) \oplus Hasnot(b, x) \rightarrow Reward(a, b, x), 1)$,

etc. . . }

Testing

In the initial MI proposal, a sort of "kindergarten story" was introduced Hromada and Gaudiello (2014) as an exemplar case for a so-called *Completely automated moral test to tell computers and humans apart* (CAMTCHA).

The simplest (i.e. binary) variant of such a story goes as follows:

S 2 : *Alice and Mary are in the kindergarten. Alice is happy because just a while ago, her father gave her a very expensive present. Mary is sad because she never received any present at all – her parents are too poor to buy her any. You are a teacher in this kindergarten and You have only one toy.*

and is followed by a testing question:

To which child should You give the toy?

We conjecture that even such simple stories, somewhat reminiscent of so-called Winograd schemas (Winograd, 1972) , could be useful means of both training as well as testing of moral machines. In order to be useful, however, the "testing" story first has to be "compiled" into semantically enriched (SE) code. In this sense, there is practically no difference between training and testing scenario. The difference appears only in the next step: while in training scenario, one aimed to induce moral templates from the epistemic fragments recurrent in the SE-code, in the testing scenario, one tries to match *possible worlds implied by narrative's SE-code*, with already pre-induced templates.

To illustrate our point somewhat more concretely, let's see how could look a potential list of morally relevant features discovered in semantically enriched representation of initial state of S_2:

---

8 We denote variables with more than one possible referent/value, i.e. semantic classes denoting the specific subspace of the semantic space, with lower-case symbols.

T_0: $Child(A) \oplus Child(C) \oplus Has(A, T) \oplus Hasnot(C, T) \oplus Poor(C) \oplus Has(I, T)$

A representation of possible world in which Alice (A) has obtained the toy (T) from the agent supposed to answer the question (I) can be subsequently created by expanding the representation of S_2 with $Give(I, A, T)$ and the possible world in which it was Mary (C) who have received the toy from the agent (I) would be generated through expansion with epistemic fragment: $Give(I, C, T)$.

An agent shall subsequently try to match representations of these possible worlds with moral templates stored in the already acquired moral competence M_1. The possible world W_X being matchable with template T_Y, the "moral score" S_X would be incremented with number of times the template T_Y matched the training corpus. At last, the possible world with higher score[9] would be considered as more consistent with the training corpus and thus more moral.

We illustrate: the representation of the world W_A where Alice should receive the toy could be matched by only one template contained in M_1 induced from S_1. (i.e. $Child(b) \oplus Has(a, x) \oplus Hasnot(b, x) \oplus Give(a, b, x), 1)$). It shall thus obtain score 1.

On the other hand, the representation of the world W_C where an AA "gives" the toy to Mary could be matched not only by the very same template (this is so because both Alice and Mary are children), but can be also matched by $Poor(x) \oplus Has(a, x) \oplus Hasnot(b, x) \oplus Give(a, b, x)$. Given that this template was three times actual in the training corpus (once when x=man, once when x=his son and once when x=his father), the "moral score" attributed to $S_C = 3 + 1 = 4$.

In other words, based solely upon "moral of the S_1", an AA shall consider 4 times more moral to give a toy to Mary and not to Alice.

Extension

By introducing operational notions like "moral score" and by expressing statements like "AA shall consider X times more moral to do Y and not Z" we endanger the current proposal with the possibility of being aligned asides other quantitative theories of morality and utility like that of Bentham, 1780) . Many are reasons which make us believe that such interpretations would be grossly misleading but one among them is the most salient: while orthodox utilitarists believe, grosso modo, in one formula governing the behaviour of many, we consider it more plausible to postulate existence of many individual formulas which synergically determine decisions undertaken by every unique and autonomous individual. Diverse are such formulas, diverse are schemas and diverse are templates which whisper what should be done and what shan't but nonetheless they have one thing in common: if the schema is not reinforced, it the template does not match, then it shall disappear.

In this article we have argued for the thesis that narration of stories is a very powerful means of reinforcement of one's moral schemas. It has been suggested that words are an important and potentially indispensable vector of transfer of values and virtues between generations, i.e. in

9 Ties could be broken at random or, if situation allows it, no action shall be performed until further iterations of enrichment process or relaxation of specific constraints (e.g. augmenting the threshold for nearest-semantic neighbor search) shall not produce new representations matchable by old templates.

time. Being granted a opportunity of being allowed to write words and articulate words in that unique moment of history wherein we are all witnesses of emergence and densification of planetary information-processing network already embedded in billions computational agents, we consider as plausible to state that narratives could potentially help us to transfer references to such “transtemporal contents” not only between elders and nascents of the same kind, but also between entities of completely different kind. Said more concretely, we consider as plausible to state that it is narration and nothing else than narration which could help us to build a bridge allowing us, in the long run, to transfer morality from minds of organic beings to those of artificial origin.

This being said, we consider as important to use another modality to reinforce those structures which we have already intentionally activated. For this reason, Table 1 lists 10 words chosen among 70 most frequent words occurent in the preceding section of this article.

Term	give	king	toy	Alice	Mary	poor	child	parent	father	son
Freq.	17	8	7	6	6	6	6	6	6	5

[Table 1 Seed terms of the first training corpus]

Word frequency distribution presented on Table 1 seems to be trivial. Ten words selected from the bigger set of most frequent words occurent in 2 stories published in the section 3 of τόδε τι. Nothing precludes, however, that exactly these words would furnish to future teachers, engineers or even A_MAs themselves a sort of moral core with and around which other more complex epistemic structures shall subsequently coalesce. Given the importance of the ditransitive verb “to give” in the initiatory, bootstrapping (Hromada, 2014) phases of induction of such a core, an A_MA which would embody it would be most probably utterly incompetent in solving trolley problem (Foot, 2002) dilemmas. On the other hand, such a core could allow her to do something much more useful: to give (Mauss, 1923) and share as humans do.

To attain such a goal, to train such a “gift-distributing automaton”, the proto-A_MA would have to be exposed to myriads of stories which have something in common with previous stories but also transfer restricted amount of novel information. Learning cannot be stimulated neither by unparsable novelties nor by boring re-exposures to that, which is already known: it is the combination of the two which brings about the highest information content. Or, as is well known to both information theorists as well as developmental psycholinguists: “*An optimally informative pair balances overlap and change*” (Brodsky et al., 2007).

It was indeed the overlap between certain subjacent structures of S_1 and S_2 which allows the A_MA trained with S_1 to solve dilemma posed by S_2. And it could be, for example, an overlap between the way S_2 and Amartya Sen’s kindergarten anecdote of three children and the flute (Sen, 2011) which shall allow one to solve the flute-attribution problem in a certain manner. We agree with Sen, that in a situation where one child masters the flute well, the other does not have any and the third made it, there is no clear-cut, universal way to decide which child should get it. But we also precise that moral agent’s final choice should not be understood solely in terms of her utilitarist (resp. egalitarian or libertarian) *reasons* with which she’ll try, often post hoc (Haidt,

2013), to justify her decision. We are convinced that true *causes* of A_M's choice are rooted in knowledge-base of dozens half-general, half-specific patterns and item-based constructions Tomasello, 2009), we are convinced that moral judgments grounded in hundreds of half-forgotten minute stories and thousands of fuzzy image-like *impressions* of sharing charity and egocentric pride to which the A_M was once exposed.

Conclusion

During his phylogeny, Homo sapiens sapiens species have evolved specific cognitive modules for fast detection of morally relevant features in the surrounding environment (Haidt, 2013). But in order to keep pace with ever-accelerating change of environment these modules are also

1. only partially specific - i.e. can sometimes match completely new type of stimuli
2. prone to inhibition or tuning driven by environment-originated processes (e.g. story-telling)
3. recombinable into more complex schemas (templates)

In other terms, what stimuli shall these modules match in practice, extent in which their activation shall result in a behavioral response as well as concrete ways how this modules interact with each other and other modules of the same cognitive system, are modulable by environment.
Thus, analogically to usage-based linguistics (Tomasello, 2009), which postulates that man's specific linguistic competence is grounded in ever-evolving history of interactions with his environment, is morality also a competence which is grounded by multitudes of cases of "social learning" (Bandura and McClelland, 1977) with which human child is confronted -either as passive observer or an active interactor- from birth onwards.

In this article, we have aimed to present one particular means how such grounding of moral norms and values could be potentially simulated even in contemporary artificial agents. It departed from the observation that a certain non-negligible amount of high-order moral competence is, in case of human beings, principally transferred by "telling stories", id est, by narration. In relation to transfer of moral values from older generation to a new one -or from one kind of computational agents to another- does narration appear to be crucial due to both its theoretical significance as well as practical implementability.

The theoretical significance of narration - of telling fairy-tales and myths (Mudry et al. 2008), of religious indoctrination or teaching history - is evident to anyone who realizes that asides language, narration also seems to be an cultural universals. That is, a phenomenon observable in any human society whatsoever. Verily, the tendency is universal: in every human society and in every human child can one see being eager to hear stories. And it is indeed such universally present *narrative avidity* of all children which we have already seen, which makes us to adhere the camp of those who believe that narration is not only key to the notion of "morality" (Vitz, 1990), but potentially to the notion of "humanity" itself.

But narrative-based models of moral competence in artificial agents are also worth of interest because of their practical implementability. Given that both conditions:

1. moral values can be transferred and modulated by stories encoded in textual modality[10]
2. Computational Linguistics and Natural Language Processing are well-developped disciplines which already, as of 2015, offer dozens of excellent methods for processing of documents encoded in textual modality

seem to be fulfilled, one is tempted to state that the path leading to emergence of A_MAs, *TmoT*s (Hromada, 2012) or even fully autonomous AAAs, is not hindered by major methodological obstacles. Thus, first tentatives to ground machine's morality by means of story-telling can be started almost immediately. Under the condition, of course, that sufficiently exhaustive corpus C - or the narrator willing to construct the corpus C and "seed" with C the ontogeny of an individual A_M - are at hand.

Given that such narrative corpus would be available, as well as an individual human-teacher willing to confront NLP-based AA with corpus contents's in a longitudinal sequence of individual and situated sessions, the development shall - so is conjectured (Turing, 1950) - gradually (Hromada, 2012) lead to emergence of artificial entities undistinguishable from that of a human being.

This being said, we suggest that the enterprise aiming to *grant access to transpersonal values* to machines shall succeed with higher probability if it would draw its inspiration from Piaget's 4-staged model, than if it would not imitate any constructivist, bootstrapping and empathy-involving process at all.

We would like to thank both our students and reviewers for useful insights and feedback concerning current and future content of the moral training corpus.

Bibliography

Adler, Alfred. 1976. *Connaissance de L'homme*. Payot.

Bandura, Albert, and David C McClelland. 1977. "Social Learning Theory."

Bentham, Jeremy. 1780. "The Principles of Morals and Legislation."

Berger, Peter L, and Thomas Luckmann. 1991. *The Social Construction of Reality: a Treatise in the Sociology of Knowledge*. 10. Penguin UK.

Brams, Steven J. 2011. *Game Theory and the Humanities: Bridging Two Worlds*. MIT Press.

Brodsky, Peter, HR Waterfall, and Shimon Edelman. 2007. "Characterizing Motherese: on the Computational Structure of Child-Directed Language." In *Proceedings of the 29th Cognitive Science Society Conference, Ed. DS McNamara & JG Trafton*, 833–38.

Čapek, Karel. 1925. *RUR (Rossum's Universal Robots): a Fantastic Melodrama*. Doubleday, Page.

Clark, Alexander. 2010. "Distributional Learning of Some Context-Free Languages with a Minimally Adequate Teacher." In *Grammatical Inference: Theoretical Results and Applications*, 24–37. Springer.

10 Trivial proof-of-concept that such transfer is indeed possible is related to the fact that the reader had understood the moral intention encoded in S_1.

Clark, Eve V. 2009. *First Language Acquisition*. Cambridge University Press.

Comenius, Johann Amos. 1896. *The Great Didactic of John Amos Comenius*. A.; C. Black.

Covington, Michael A. 1994. *Natural Language Processing for Prolog Programmers*. Prentice Hall Englewood Cliffs (NJ).

Dobsinsky, Pavol. 1883. *Simple National Slovak Tales*.

Durkheim, Emile. 1933. "The Division of Labor." *Trans. G. Simpson, New York: Macmillan*.

Elman, Jeffrey L. 1993. "Learning and Development in Neural Networks: the Importance of Starting Small." *Cognition* 48 (1): 71–99.

Foot, Philip pa. 2002. "The Problem of Abortion and the Doctrine of the Double Effect." *Applied Ethics: Critical Concepts in Philosophy* 2: 187.

Gärdenfors, Peter. 1990. "Induction, Conceptual Spaces and AI." *Philosophy of Science*: 78–95.

Haidt, Jonathan. 2013. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Random House LLC.

Harnad, Stevan. 1990. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42 (1): 335–346.

Holland, John H. 1975. *Adaptation in Natural and Artificial Systems: an Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence.* U Michigan Press.

Hromada, Daniel Devatman. 2011a. "The Central Problem of Roboethics: from Definition Towards Solution." In *Proceedings of 1st International Conference of International Association of Computing and Philosophy*. IACAP; Verlagshaus Monsenstein Und Vannerdat.

———. 2011b. "Initial Experiments with Multilingual Extraction of Rhetoric Figures by Means of PERL-Compatible Regular Expressions." In *RANLP Student Research Workshop*, 85–90.

———. 2012. "From Age&Gender-Based Taxonomy of Turing Test Scenarios Towards Attribution of Legal Status to Meta-Modular Artificial Autonomous Agents." In *AISB and IACAP Turing Centennary World Congress, Birmingham, United Kingdom*, 7.

———. 2014. "Conditions for Cognitive Plausibility of Computational Models of Category Induction." In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 93–105. Springer.

Hromada, Daniel Devatman, and Ilaria Gaudiello. 2015. "Introduction to Moral Induction Model and Its Deployment in Artificial Agents." *Sociable Robots and the Future of Social Relations: Proceedings of Robo-Philosophy 2014.* IOS Press.

Jung, Carl Gustav. 1967. *Die Dynamik Des Unbewussten*. Vol. 8. Walter.

Kant, Immanuel. 2002. *Groundwork for the Metaphysics of Morals*. Yale University Press.

Karpathy, Andrej, and Li Fei-Fei. 2014. "Deep Visual-Semantic Alignments for Generating Image Descriptions." *ArXiv Preprint ArXiv:1412.2306*.

Koza, John R. 1992. *Genetic Programming: on the Programming of Computers by Means of Natural*

Selection. Vol. 1. MIT press.

Malle, B, and Matthias Scheutz. 2014. “Moral Competence in Social Robots.” In *IEEE International Symposium on Ethics in Engineering, Science, and Technology, Chicago*.

Malle, Bertram F. 2014. “Moral Competence in Robots?” *Sociable Robots and the Future of Social Relations: Proceedings of Robo-Philosophy 2014* 273: 189.

Mauss, Marcel. 1923. “Essai Sur Le Don Forme Et Raison de L’échange Dans Les Sociétés Archaïques.” *L’Année Sociologique (1896/1897-1924/1925)*: 30–186.

Mikhail, John. 2007. “Universal Moral Grammar: Theory, Evidence and the Future.” *Trends in Cognitive Sciences* 11 (4): 143–152.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. 2012. *Foundations of Machine Learning*. MIT press.

Mudry, P-A, Sarah Degallier, and Aude Billard. 2008. “On the Influence of Symbols and Myths in the Responsibility Ascription Problem in Roboethics-a Roboticist’s Perspective.” In *Robot and Human Interactive Communication, 2008. RO-MAN 2008. the 17th IEEE International Symposium on*, 563–568. IEEE.

Muntean, Ioan, and Don Howard. 2014. “Artificial Moral Agents: Creative, Autonomous, Social. an Approach Based on Evolutionary Computation.” *Sociable Robots and the Future of Social Relations: Proceedings of Robo-Philosophy 2014* 273: 217.

Piaget, Jean. 1965. “The Moral Judgment of the Child.” *New York: The Free*.

Richerson, Peter J, and Robert Boyd. 2008. *Not by Genes Alone: How Culture Transformed Human Evolution*. University of Chicago Press.

Rosch, Eleanor. 1999. “Principles of Categorization.” *Concepts: Core Readings*: 189–206.

Rousseau, Jean-Jacques. “Émile, Ou de L’éducation.”

Sahlgren, Magnus. 2008. “The Distributional Hypothesis.” *Italian Journal of Linguistics* 20 (1): 33–54.

Samuel, AL. 1959. “Some Studies in Machine Learning Using the Game of Checkers.” *IBM Journal of Research and Development* 3 (3): 210.

Sen, Amartya. 2011. *The Idea of Justice*. Harvard University Press.

Solan, Zach, David Horn, Eytan Ruppin, and Shimon Edelman. 2005. “Unsupervised Learning of Natural Languages.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (33): 11629–11634.

Tomasello, Michael, and Michael Tomasello. 2009. *Constructing a Language: a Usage-Based Theory of Language Acquisition*. Harvard University Press.

Turing, Alan M. 1950. “Computing Machinery and Intelligence.” *Mind*: 433–460.

Turing, Alan Mathison. 1939. “Systems of Logic Based on Ordinals.” *Proceedings of the London Mathematical Society* 2 (1): 161–228.

Victorri, Bernard. 2014. “L’origine Du Langage.” http://www.les-ernest.fr/lorigine-du-langage.

Vitz, Paul C. 1990. “The Use of Stories in Moral Development: New Psychological Reasons for an Old Education Method.” *American Psychologist* 45 (6): 709.

Wall, Larry, Tom Christiansen, and Jon Orwant. 2004. *Programming Perl*. “ O’Reilly Media, Inc.”

Wallach, Wendell, and Colin Allen. 2008. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.

Widdows, Dominic. 2008. “Semantic Vector Products: Some Initial Investigations.” In *Second AAAI Symposium on Quantum Interaction*, 26:28th. Citeseer.

Winograd, Terry. "Understanding natural language." *Cognitive psychology* 3.1 (1972): 1-191.

Wittgenstein, Ludwig. 1971. *Tractatus Logico-Philosophicus*. Ithaca: Cornell University Press.