DOCTORAL THESIS

A Study on Implementing of Culture, Religion and Time-Awareness to Machine Ethics Algorithms

by

Jagna Nieuważny

Advised by:

Fumito Masui

Michal Ptaszynski

KITAMI INSTITUTE OF TECHNOLOGY
GRADUATE SCHOOL OF ENGINEERING

March 2021

Acknowledgements

This work would be not be possible without the support that I received from many people. I would like to reserve this page to thank them.

I want to thank my supervisors, Professor Fumito Masui and Professor Michal Ptaszynski, for their time, kindness, understanding, guidance and feedback throughout this project.

The Kitami Institute of Technology gracefully agreed to exempt me from fees for my doctoral course, making it easier for me to concentrate on research.

My patient husband Karol has always been ready to give me encouragement, advice and invaluable support in all things technical and my family kept nagging me to get a PhD until I finally did. I am grateful to each and everyone of you.

Jagna Nieuważny

March 2021

ABSTRACT

Recent rapid developments in the field of Artificial Intelligence (AI), especially those regarding the development and implementation of artificial neural networks, surface the questions previously considered to arise much later in the future and were so far only addressed in philosophy, or its lighter derivatives, such as science fiction literature. One of such questions regards the limitations of Artificial Intelligence systems to fully grasp the importance and necessity in human lives of such subjective, spiritual and motivational phenomena as religious experience or a moral outlook. In this dissertation, I present the results of my research devoted to applying Artificial Intelligence to quantitative analysis of factors influencing the ethical outlook of Japanese people.

I begin with presenting the background of this research. In particular, I focus on the relations between Artificial Intelligence and ethics, seen as a set of moral rules transmitted through religion, education or historical experience.

After that, I report the results of experiments conducted in the course of my research. As an introduction to the first experiment, I provide an overview of previous cross-sectional research in the field of Religion and Technology. I then analyze how much religious vocabulary, in particular Buddhist vocabulary taken from the largest online dictionary of Buddhist terms, is present in everyday social space of Japanese people, particularly, in Japanese blog entries appearing on a popular blog service (Ameba blogs). I interpreted the level of everyday usage of Buddhist terms as appearance of such terms in the consciousness of people. I further analyzed what emotional and moral associations such contents generate. In particular, I analyzed whether expressions containing Buddhist vocabulary are considered appropriate or not from a moral point of view, as well as the emotional response of Internet users to Buddhist terminology.

As a result of analyzing the data, I found out that Buddhist terms were in fact not absent as a theme from Japanese blogs and generated a strong emotional response. However, while the general reaction to several expressions using Buddhist terms was as expected, there were sometimes surprising twists in terms of social consequences, major discrepancies between what

is perceived as ethically correct behavior between the Buddhist doctrine and the reasoning of the general population, as well as a considerate number of terms which have lost their original meaning and instead became slang expressions.

Secondly, I focus on ethical education as a means to improve artificial companion's conceptualization of moral decision-making process in human users. In particular, I focus on automatically determining whether changes in ethical education influenced core moral values in humans throughout the century. I analyze ethics as taught in Japan before WWII and today to verify how much the pre-WWII moral attitudes have in common with those of contemporary Japanese, to what degree what is taught as ethics in school overlaps with the general population's understanding of ethics, as well as to verify whether a major reform of the guidelines for teaching the school subject of "ethics" at school after 1946 has changed the way common people approach core moral questions (such as those concerning the sacredness of human life). I selected textbooks used in teaching ethics at school from between 1935 and 1937, and those used in junior high schools today (2019) and analyzed what emotional and moral associations such contents generated. The analysis was performed with an automatic moral and emotional reasoning agent and based on the largest available text corpus in Japanese as well as on the resources of a Japanese digital library.

As a result, I found out that, despite changes in stereotypical view on Japan's moral sentiments, especially due to historical events, past and contemporary Japanese share a similar moral evaluation of certain basic moral concepts, although there is a large discrepancy between how they perceive some actions to be beneficial to the society as a whole while at the same time being inconclusive when it comes to assessing the same action's outcome on the individual performing them and in terms of emotional consequences. Some ethical categories, assessed positively before the war, while being associated with a nationalistic trend in education have also disappeared from the scope of interest of post-war society.

Finally, in the course of a third experiment and based on the findings of the two previous ones, I try to answer a twofold question. Firstly, since the methods used for performing authorship analysis imply that an author can be recognized by the content he or she creates, I was interested in finding out whether it would be possible for an author identification solution to correctly attribute works to authors if in the course of years they have undergone a major psychological transition. Secondly – and from the point of view of the evolution of an author's ethical values – I checked

what it would mean if the authorship attribution system encounters difficulties in detecting single authorship, hypothesizing that it could mean that historical events had had significant impact on the person's ethical outlook. I set out to answer those questions through performing a binary authorship analysis task using a text classifier based on a pre-trained transformer model.

As a result, I was able to confirm that in the case of texts authored by my target author, Arata Osada, in a time span of more than 10 years, while the classification accuracy drops by a large margin and is substantially lower than for texts by other non-fiction writers, the confidence of the model in its predictions remains at a similar level as in the case of a shorter time span, which means that the classifier was in many instances tricked into deciding that texts written by Arata Osada over a time span of multiple years were actually written by two different people, which in turn leads me to believe that a such a change can affect authorship analysis and historical events have great impact on a person's ethical outlook as expressed in their writings.

Based on the findings from those experiments, I outline future research goals.

ABSTRACT IN JAPANESE (論文内容の要旨)

近年における人工知能（AI）分野の急速な発展あ目覚ましく，ニューラルネットワークの実用化は，これまで近い将来には実現は可能と考えられてきた問題，例えば，哲学やサイエンス・フィクションでしか扱われて来なかった問題を議論の遡上に引き上た．そのひとつに，人工知能に宗教的経験や道徳的観念を把握させる問題がある．これは，日常生活における主観的、精神的動機付けとなりうる人間の倫理観を人工知能で扱おうとする問題である．本論文では，日本人の倫理観に影響を与える要因を定量的に分析するために、人工知能技術を活用することに注目した一連の研究成果を報告する．

まず，本研究の背景を紹介し，人工知能と宗教，教育，或いは歴史的経験を通して共有される一連の道徳規範としての倫理観について分析する． 次に，研究の過程で行った実験とその結果について述べる．実験への導入として，宗教と技術の分野におけるこれまでの横断的研究の概要を説明する．そして，大規模日本語コーパス（YACISコーパス）に現れる仏教用語の出現率や仏教に基づいた倫理概念の有無，またはそれらとの感情的，道徳的関連性を分析し，現代日本人が持つ仏教教義に対する意識レベルの定量化を試みる． さらに，現存する最大規模のオンライン仏教用語事典を用いて，YACISコーパスから事典に登録された用語を含む文章を抽出，分析する．加えて，倫理推論エージェント（Moral Reasoning Agent）を抽出文に適用することで，文が持つ道徳的概念の把握と感情要素や倫理的要素の分析を行った．その結果，YACISコーパスには，仏教に基づいた倫理概念が存在していることはもちろん，現代日本人は倫理的に正しいとみなされる行動を正しい行動として意識しているが，場合によっては正しい行動を苦しいと捉えるケースもあることを明らかにした．

以上を踏まえ，人間の持つ道徳的概念に対する倫理教育の影響を考察する．日本において，「道徳」を義務教育科目とするガイドラインの大幅改訂により，一般人の道徳の捉え方がどの程度変化したかを検証する．具体的には，①第二次世界大戦以前の「修身」と②現代日本の教育制度における「道徳」の教科書内容を比較分析し，前者の道徳概念が現代日本人の倫理観とどの程度共通しているか，または，教育において倫理的に正しいとみなされる行動が一般的な倫理の理解とどの程度共有されているかについて検証する． まず，1930年代の

高校授業の目録である「修身教授録」（戦前データセット）と現代の中学授業で使用されている「私たちの道徳」（戦後データセット）から，『人生」または「命」に関連するフレーズを抽出する．次に，青空文庫収録の1930年代出版小説を対象に，それらの小説中に出現する「命」に関連するフレーズを抽出し，戦後データセット中に出現するフレーズと比較する．さらに，倫理推論エージェントを用いて，それらの内容の感情的，道徳的関連性を分析する． 分析の結果，教育制度の重大な改革にも関わらず，日本人の根本的倫理観には変化が見られず，過去においても現代においても日本人は特定の基本的道徳概念を共有していることが明らかになった． そして，日本人は，道徳的とみなされる行動が社会全体にとって有益であることを認識しながらも，それを実行する個人に対しては必ずしも楽観的結果をもたらさないことも認識していること，すなわち，「個人の利益」と「社会の利益」の相違が存在することも明らかとなった．

最後に，日本人が持つ倫理観に影響を与える要因を吟味する．歴史的事象の影響を受けたとされる複数の著述家による著作物を詳細に分析し，道徳観の変化を定量化することを試みる．具体的には，以下の二つの問に答えるものである． 第一に，既存の著者属性認識技術を用いて，長年に渡って様々な要因（時間経過，経験，意見の変化など）によって執筆スタイルが変化した同一著者の作品を正しく認識できるかどうかを検証する． 第二に，同一著者判定が困難なケースに注目し，その原因を調査することで，歴史的事象が人間の倫理観に重大な影響を与える可能性があるという仮説を立て，これを検証する． 上記の問に答えるために，著者属性認識タスクを実行した．訓練データとして大規模日本文学コーパスである青空文庫コーパスを用い，事前学習によるトランスフォーマモデルを用いて著者属性判定器を構築した．テストデータには，太平洋戦争前後において大きく主張を変えた著名な教育史学者，長田新の著作物を用いた． 実験の結果，時間経過，経験，意見記述の変化が著者属性認識性能に影響を与えることが確認された．このことは，歴史的事象が同一著者の倫理観に大きな影響を与えることを示唆するものである．

以上そ総括すると，今後設計される人工知能システムは，倫理観を一定のものとするのではなく，道徳的推論における歴史的，宗教的，文化的、時間の経過がもたらした変化の影響を考慮したものとする必要があるという結論が得られる．

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Since moral and ethical education is an indispensable element in the formation of the character of any human being, any AI system whose task is to accompany humans (artificial companions, etc.) is by definition obliged to have an ethical insight as well. However, reducing complex ethical and moral questions to a set of universal rules or core moral values poses a big challenge to humans, let alone today's AI systems.

While the past years have brought an exponential growth in the development of practical solutions in the field of Artificial Intelligence, not enough attention has been given to more complex issues related to equipping newly created solutions with an ethical outlook or a moral code.

However, exploring today how technology is predictable to develop in the areas of ethics and understanding of morality, and what influences people's moral choices and preferences, is important for appropriate appraisal and appreciation prior to having the developments already materialize, and then scrambling to react after the fact.

I strongly believe that the field of Natural Language Processing (NLP) has an important role to play in the urgent tasks of providing machines with an ethical core. This dissertation sums up my research performed in the course of my doctoral thesis and aimed at contributing to the evolution of machine ethics through studying how to implement Culture, Religion and Time-Awareness to Machine Ethics Algorithms.

1.1 Structure of the Thesis

The remainder of this thesis is organized as follows. Chapter 2 describes the background of my research. In particular, I review some of the related research in the area of AI and religion as well as AI and ethics.

In Chapter 3, I provide an overview of some of the previous cross-sectional studies in the field of Religion and Technology. After introducing the main ressources and tools used in performing this research, I then report the outline and results of a preliminary experiment aimed at analyzing how much religious vocabulary, in particular Buddhist vocabulary taken from the largest online dictionary of Buddhist terms, is present in everyday social space of Japanese people, particularly, in Japanese blog entries appearing on a popular blog service (Ameba blogs).

In Chapter 4, I try to automatically determine whether changes in ethical education influenced core moral values in humans throughout the century. I analyze ethics as taught in Japan before WWII and today to verify how much the pre-WWII moral attitudes have in common with those of contemporary Japanese, to what degree what is taught as ethics in school overlaps with the general population's understanding of ethics, as well as to verify whether a major reform of the guidelines for teaching the school subject of "ethics" at school after 1946 has changed the

way common people approach core moral questions (such as those concerning the sacredness of human life).

In Chapter 5, I present my aim at finding out whether a model created for the task of same authorship detection would correctly attribute works from different points in time to the same author and with what accuracy, in order to see if the system experiences any additional difficulty in single authorship identification when presented with two texts by a person whose opinions and/or ethical values changed in the intervening period between writing – which would mean that the impact of historical events on a person's ethical outlook and the content (books, articles, etc.) he or she produces, is significant enough that it can be quantified.

Chapter 6 concludes the thesis, with a review of the principal findings of this research and ideas for future work.

Chapter 2

AI and Ethics – Background

2.1 The evolution and philosophical foundation of Machine Ethics

One of the milestones often considered to be a starting point in the discussion about Artificial Intelligence is the concept of an AI system obtaining a singularity (a self-conscious state). Kurzweil [30] came up with a prediction that the year when artificial intelligence would exceed human intelligence and thus achieve singularity would be 2045. Although there has been a debate concerning the accuracy of Kurzweil's vision as a whole, as well as of the predicted time frame, the developments within the field of AI every year provide new points of view in the discussion on whether or not the singularity would occur and when it could most probably happen. However, most of the scientific community agrees that if it was to occur, one must focus on private and personal aspects of user experience [20], which includes both emotions as well as more culture-related areas, such as beliefs, moral code, or religious thought and spirituality.

This goes along with the especially recently noticeable new rising trend in the development of technology, namely to create technology personalized so that it is capable of responding to the private needs of a diversified range of users. Therefore, the designers and programmers of future technology cannot ignore the important social factor that is the religiousness and ethical code (or lack of it) among its users and thus an intelligent technology should also be aware of the socio-cultural impact of religion and ethics. As such, whoever creates the technology, needs to take into account the religious psyche and ethical outlook of the user and be aware of the fact that adherents to different cultures, and thus also different religions and different ethical codes have different spiritual and moral needs and demands towards technology.

There is no one specific "ideal religion" or "model ethic" in the relationship between Artificial Intelligence and societies that would be perfectly suited to be implemented as a robot's base for a set of universal moral or ethical principles. However, as most religious doctrines or ethical systems include also specified sets of moral rules, personalized technology interacting with human users needs to be aware of the religious systems of such users and thus also be able to recognize and process moral rules of the specified religion or a specific moral code.

A sub-field of AI recently gaining in popularity focused on designing such models of moral rules and ethical systems is called Machine Ethics, sometimes referred to as Machine Morality[1], Computational Ethics[2] or Computational Morality [57]. Unlike the field of Computer Ethics – which has traditionally focused on ethical issues surrounding humans' use of machines – Machine Ethics is concerned strictly

[1] http://www.yalescientific.org/2012/05/machine-morality-computing-right-and-wrong/

[2] http//:www.demo.clab.cs.cmu.edu/ethical_nlp

with ensuring that the behavior of machines toward human users, and perhaps other machines as well, is ethically acceptable.

The beginnings of Machine Ethics as a practical-experimental science field can be traced back to Rzepka and Araki [59], who proposed an extension to their Web-based knowledge discovery system GENTA (General Belief Retrieving Agent) that searches the Web for opinions, typical behaviors, common consequences and exceptions, by counting ethically relevant neighboring words and phrases, aligning these along a continuum from positive to negative behaviors, and subjecting this information to statistical analysis, which would lead to the development of a majority-rule ethics (or "morality of Web-crowds", as called later in [28]), useful in guiding the behavior of autonomous systems.

In 2006, Oxford University Press published Moral Machines, Teaching Robots Right from Wrong, advertised as "the first book to examine the challenge of building artificial moral agents, probing deeply into the nature of human decision making and ethics" ([67]).

In 2011, Cambridge University Press published a collection of essays about machine ethics edited by Michael and Susan Leigh Anderson ([3], who also edited a special issue of IEEE Intelligent Systems on the topic in 2006 ([4]). This special issue contained essays by, among others, James Moor, who defined five possible ways in which values could be ascribed to machines, while Bruce McLaren argued that he was reluctant to give machines the power to make ethical decisions by themselves, preferring instead to have machines, in an advisory capacity, inform human users of solutions to previous cases similar to the dilemma, without reaching a decision autonomously.

In 2014, the US Office of Naval Research[3] announced that it would distribute 7.5 million USD in grants over five years to university researchers to study questions of machine ethics as applied to autonomous robots . Among other governmental initiatives, the AI Ethics Committee was initiated in 2014 by the Japanese Society for Artificial Intelligence[4] to discuss some ethical implications of the singularity in Artificial Intelligence and in 2016 the European Parliament published a paper[5] to encourage the Commission to address the issue of robots' legal status.

Even if we agree that there should be ethical principles at play when creating artificial intelligence systems, there exists an unsolved discussion about how to implement those principles in robots (namely whether to choose an "implicit ethical agent" – a machine that has been programmed to behave ethically without an explicit representation of ethical principles and an "explicit ethical agent", able to calculate the best action in ethical dilemmas, not restrained by the ethics of its creator) and which exactly ethical principles to choose [27].

There is also a more profound problem with systems attempting to perform moral judgements. Namely, a machine that has learned, or was programmed, to make correct ethical judgments (even if it does it perfectly), but does not have the principles to which it can appeal to justify or explain its judgments, is lacking an essential component to being accepted as an ethical agent[6] – this point was first brought up by Immanuel Kant, who made a distinction between an agent that acts from a sense of duty (consciously following an ethical principle), rather than merely in accordance with duty, out of which he only considered the former to be acting

[3] https://www.defenseone.com/technology/2014/05/now-military-going-build-robots-have-morals/84325/

[4] http://ffj.ehess.fr/upload/Discussion/CEAFJPDP-18-02.pdf

[5] https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html

[6] https://www.aaai.org/ojs/index.php/aimagazine/article/view/2065

in an ethical manner [25].

Some researches believe that programming ethical behavior to machines is an impossible task as ethics is not a value that could be computed. Nevertheless, already since the 19th century, attempts have been made by English philosophers, such as Jeremy Bentham [10] and John Stuart Mill to make ethics objectively measurable through means of performing "moral arithmetic". The doctrine, known as Hedonistic Act Utilitarianism, formulated in opposition to ethics based on subjective intuition, holds that the right action is the one likely to result in the greatest "overall pleasure", calculated by adding up units of pleasure and subtracting units of displeasure experienced by all those affected[7].

In their essays, Michael and Susan Anderson [3] state that there are also those doubting whether machines will ever be able of making ethical decisions since they lack emotions and thus cannot predict the feelings of those affected by their actions. This can however become an actual advantage, since there is no risk for a machine to succumb to its impulses or get carried away by emotions. Interaction with an Artificial Intelligence system can teach theorists of ethics "pure" ethical decision making, without the influence of emotions. However, research in machine ethics could also advance the study of ethical theory, with ethics being one of the most practical branches of philosophy. Research in machine ethics has the potential to discover problems with current ethical theories, perhaps even leading to the development of better ethics, as AI researchers force scrutiny of the details involved in actually applying an ethical theory to particular cases.

[7] https://phare.univparis1.fr/fileadmin/PHARE/Seminairephiloeco/V._Bianchini_-_JM___JSM_on_the_Felicific_Calculus___.pdf

2.2 A few thoughts on how technology is predictable to develop in the areas of spirituality

Recent rapid developments in the field of Artificial Intelligence (AI), especially those regarding the development and implementation of artificial neural networks, surface the questions previously considered to arise much later in the future and that were so far only addressed in philosophy, or its lighter derivatives, such as science fiction literature. One of such questions regards the limitations of Artificial Intelligence systems to fully grasp the importance and necessity in human lives of such subjective, spiritual and motivational phenomena as religious experience and the existence of an ethical compass.

Although the question of the implementation of spirituality or understanding of religious experience in machines today seems far-fetched still, it has not been long since Marvin Minsky stated that the implementation of emotional understanding in machines is irrelevant [38]. However, merely ten years later the field of Affective Computing [51], has changed the whole field of Artificial Intelligence, giving birth to presently some of the most popular research areas in AI, such as Affect Analysis or Sentiment Analysis.

Moreover, the history of technological developments in general, giving as one example the exploitation of nuclear energy, provides clear hints on the results of an irresponsible push for the fast development of new technologies, which is often referred to as the "how" of technology [13]. On the other hand, the questions of what global changes will those developments bring, whether they are anticipated, or

even necessary (the "why" question), and whether such developments even should be made (the "if" question), have been notoriously disregarded.

Therefore, exploring today how technology is predictable to develop in the areas of spirituality and understanding of religious or ethical thought, and how religious traditions or moral codes might respond and adapt to such developments, is important for the appropriate appraisal and appreciation prior to having the developments already materialize, and then scrambling to react.

Chapter 3

Statistical Analysis of Emotional and Moral Associations with Buddhist Religious Terms Appearing on Japanese Blogs

3.1 Introduction

To address questions related to the relations between religiosity and morality, in this chapter I will present my study into how religious thought is present in everyday language use, by focusing on the presence of strictly religious terms in secular media channels, and analyzing what emotional and moral connotations they associate with most commonly.

As the area of study, I chose Japanese blogs, because the largest socially bound language resource, namely a blog corpus, has been developed particularly for this

language [54]. Although there exist other, sometimes larger corpora, they are usually of mixed domains, such as conversations, newspapers, literature, or even random Web contents [23], which could provide mixed or even contradictory signals. Moreover, since the predominant religion in Japan is the Buddhist religion, I focused especially on Buddhist terminology present in such casual social creations as blogs, to see emotional and moral associations they correlate with the most.

The remainder of this chapter is organized as follows. In Section 2 I review existing research focused on the relations between religion and modern technology. Section 3 describes my experiment analyzing the emotional responses and moral evaluations related to Buddhist terms observed in Japanese blogs. In Section 4 I discuss the results of the experiment. Finally, the chapter is concluded in Section 5.

3.2 Religion and Technology – A Cross-sectional Field Study

3.2.1 McDorman-Entezari Experiment

One of the first experiments regarding the responses to technology from a religious point of view was conducted by Karl McDorman and Steven Entezari[32]. The experiment, conducted on almost 500 college students of diversified religious backgrounds, revealed that religious fundamentalists tend to view human-like robots as being more unsettling than people with no strong religious background.

Participants in the study were first asked about nine individual traits (including religious fundamentalism or the individual's tendency to feel disturbed by reminders of death or physiology) that the researchers thought might have had something in

common with the sensitivity to the uncomfortable "creepiness" of robots mimicking human appearance and behavior.

In their evaluation, McDorman and Entezari borrowed the concept of Uncanny Valley, which first appeared in an essay by Masahiro Mori in 1970 [40]. Mori discusses the fact that the affinity for a robot typically increases as it is made to look more human. However, at some point the robot may become sufficiently realistic, while its remaining non-human features become noticeable and disturbing. Uncanny valley refers here to the feeling of eeriness or discomfort related to robots that appear almost human.

MacDorman and Entezari propose that the Uncanny Valley phenomenon can consist of both culturally-conditioned feelings — such as Christian beliefs in humans being unique and set apart from robots and the rest of creation — and biologically-rooted feelings involving fear and disgust. The study was meant to trace how those individual traits affect one's sensitivity to the Uncanny Valley (a higher sensitivity being understood as higher ratings of eeriness and lower ratings of "perceived warmth" for android robots).

Next, the study asked participants to rate a series of six videos showing five robots and one human based on factors such as eeriness and warmth. The robot samples included both robots that do not resemble humans in appearance (such as the iRobot Roomba vacuum cleaner) as well as a number of human-like androids with non-human features such as open skulls with exposed wires, expressionless faces, mechanical body movements, and voices not synchronized with lip movements.

From the results of their study MacDroman and Entezari concluded that Religious Fundamentalism might heighten the sensitivity to the Uncanny Valley by operating through related sociocultural constructions, such as the conviction

that human beings are unique – set apart from robots and the rest of natural creation. This in order would allow us to speculate that believing in such concepts as salvation or eternal life would make the Uncanny Valley sensitivity stronger especially for the Christian worldview in which human beings are meant to be created in the image of God and set apart from all of his other creations.

The proposition that fundamentalists of the Abrahamic religions (Judaism, Christianity, and Islam) are more prone than other individuals to perceive androids as cold and eerie has a twofold theoretical basis. Firstly, fundamentalists adhere to a worldview that divides humanity from the rest of results in heightened out-group derogation and negational categorization. Secondly, androids may rekindle awareness of repressed fears that are especially pernicious to fundamentalists: a "soulless" machine assuming the role of a human being renders the soul functionally superfluous. Acknowledging that a robot might own a soul and thus the creation of artificial persons would mean to those who believe in an eternal soul that what they have referred to as the soul is in fact an emergent phenomenon and a property of brain function, rather than a separate, incorporeal substance. However, this conclusion does not come without disadvantages. Firstly, it would be difficult to evaluate who exactly qualifies as a religious fundamentalist. Secondly, given that intergroup contact reduces prejudice and anxiety, those who have had more exposure to androids may also have fewer negative attitudes. For followers of any given religion prone to take into consideration the recommendations of religious leaders, they might assume it unacceptable to keep in touch with modern technology and thus be less acquainted with robots.

MacDorman and Entezari themselves also acknowledge that a broader principle may be at play, namely, a preference for things to be clearly distinguishable (e.g.,

"good vs. evil," "human vs. inhuman" etc.) which fundamentalists share with other groups, such as those rating high in right-wing authoritarianism.

Whether one agrees with the conclusion drawn from said experiment or not, it goes to prove that a different religious background will definitely influence the way an individual interacts with technology in general and Artificial Intelligence in particular. If such are observations concerning followers of so-called "book religions" (Islam, Judaism or Christianity), it might prove interesting to consider how different the approach to an Artificial Intelligence system, such as a robot with sophisticated functionalities, performing religious duties or simply being the follower of any given religion might be among followers of a religion that does not stipulate the existence of a permanent soul and does not stress the human–nonhuman distinction, as it is in Buddhism.

3.2.2 Robots from Viewpoint of Religion

In parallel to broad questions in psychology such as what constitutes a personality as far as being a human, there has been scholarly discussions concerning the design and manufacturing of sentient humanoid robots [22].

If we assume that religion is one of the elements important in the development of a personality, then it also needs to be addressed that, according to Guthrie [19], forming of religion itself in humans has its cognitive central source in anthropomorphism. Guthrie claims religion can best be understood as systematic anthropomorphism – that is, the attribution of human characteristics to nonhuman things and events – and that religion consists of seeing the world as human-like (through humanizing weather phenomena, speaking of "raging fires" and finding a deeper meaning in random events such as earthquakes).

Guthrie points out that our tendency to find human characteristics in the nonhuman world stems from a deep-seated perceptual strategy: in the face of pervasive (if mostly unconscious) uncertainty about what we see, we bet on the most meaningful interpretation we can. In scanning the world, we always look for what most concerns us – living things, and particularly, humans.

Here, especially Buddhism provides an interesting background for discussion on the role of technology, e.g., robots, in one's spiritual life. Particularly, the Buddhist term *anātman* refers to the doctrine of "non-self", meaning that there is no unchanging, permanent self, soul or essence in living beings. What appears to be a seemingly singular, permanent self or soul is actually a composition of five ever-changing elements, the *skandhas*, which together create the illusion of a fixed identity and continuous self. Moreover, it is clinging to this fixed self that creates all our everyday experiences, considered to be unnecessary suffering in this world. Here, we can imagine robots being created for the sole purpose of rescuing humans, e.g., in time of disasters[1]. Buddhist doctrine can consider a machine that revealed compassion for others, without forming attachments and without regard for its own life, as a realization of the Buddha nature in an unprecedented fashion [35].

Geraci [18] also notes different cultural attitudes and assumptions as to cultural evaluations and social acceptance of robotics between the East, especially Japan, and the so-called Western society. Specifically in the case of Japanese Buddhism, Geraci emphasizes the underlying animistic attitude toward machines among the Japanese, and points out some the similarities with concepts appearing in Western societies, such as the Greek myth of Pygmalion, creating his ideal effigy of a woman out of ivory, and the Jewish myth of Golem – molded out of clay by Judah Loew

[1] https://www.rm.is.tohoku.ac.jp/rescue+systems/

ben Bezalel, the late 16th century rabbi of Prague, and then brought to life through rituals.

3.2.3 Robots in Different Religious Industries

According to the analysis by Frey and Osborne who created the "Will robots take my job?" website[2], the chances of the responsibilities of a religious leader being automated are less than one per cent[3]. This renders religion one of the safest industries to work in for those concerned about technological unemployment. However, there have been recent attempts at employing robots in roles traditionally reserved to human religious leaders – some of which have been introduced below.

Besides engineering knowledge and design, engineers and mechanical designers are under the influence of their cultural aesthetics, and as such they design and construct culturally appropriate designs [5]. In the same manner, one could argue that robots or similar technology could be utilized for different purposes depending on cultural, political and personal contexts.

One example of this could be the invention by Akbar Rezaie, a young Koran teacher from the town of Varamin in Iran, who had been teaching children at the Alborz elementary school with the help of a humanoid robot Veldan[4]. The humanoid

[2] https://willrobotstakemyjob.com/

[3] In 2013 Carl Benedikt Frey and Michael A. Osborne published a report titled "The Future of Employment: How susceptible are jobs to computerisation?" (https://www.oxfordmartin.ox.ac.uk/publications/view/1314). It examined how susceptible jobs are to computerization, by implementing a novel methodology to estimate the probability of computerization for 702 detailed occupations, using a Gaussian process classifier. According to their estimates, about 47 percent of total US employment is at risk. Although the report is specific to the US job market, it is easy to see how this might apply all over the world. The authors of the site extracted the jobs and the probability of automation from the report and have made it easy to search for one's job while also adding some additional information from the Bureau of Labor Statistics to provide some background facts about different jobs.

[4] https://www.rt.com/news/iran-praying-robot-children-888/

robot was constructed using an educational kit called Bioloid from the Korean robot manufacturer Robotis[5]. The robot has been introduced into classes to provide a visual example of prayer that is more likely to capture the attention of children and was modified to let the robot perform praying movements, such as prostration, by modifying parts of the set, for instance adding additional servomotors.

In Pune, Western India, on the occasion of the Hindu festival of Ganesh Chartuth, where devotees gather to celebrate and perform rituals to honor the elephant-headed god, Ganesh, robotic arms designed by Patil Automation Ltd. – which mostly produces manufacturing robots – were made to perform a ritual known as aarti, where a priest moves a lamp in circles in front of the statue of a deity, while chanting hymns[6]. One robotic arm waved a small flaming pot before a small statue of Ganesh, surrounded by offerings, while the second arm rang a bell. Workers for Patil Automation insisted on clarifying that "the robot was mostly a decoration and was not intended to replace human practitioners".

In Japan, some robots have taken on religious roles. For example, a robot named Pepper developed by SoftBank Robotics has been adapted to perform traditional Buddhist funeral rites[7]. Nissei Eco Co. offers Pepper to chant Buddhist sutras at funerals, providing a cheaper alternative to human priests, charging 50,000 JPY (around $440 USD) for its services — which is about one tenth of the price to pay for a monk to provide sutra reading at a funeral). This shows a more liberal attitude of Japanese Buddhism to the implementation of robots in the religion.

A similar initiative was taken in Beijing, where a robot named Xian'er has been

[5] http://support.robotis.com/en/product/bioloid_main.htm

[6] https://eandt.theiet.org/content/articles/2017/09/robotic-arm-performs-traditional-hindu-ritual/

[7] https://eandt.theiet.org/content/articles/2017/08/pepper-the-robot-performs-traditional-buddhist-funeral-service/

serving the public by reciting sutras in English and Chinese in Longquan Monastery since 2015. Since Xian'er was created to duplicate the level of knowledge of the Buddhist doctrine similar to that of a novice monk, when asked questions that surpass its abilities it answers "I need to ask my master" or "I am just a small monk". Venerable Xianfan, a Buddhist monk and the creator of Xian'er, states that Artificial Intelligence could be used to help spread the teachings of the Buddha in China by saying that science and Buddhism are not opposing nor contradicting, and can be combined and are mutually compatible[8].

As another example from a different culture, to commemorate 500 years since Martin Luther published "The 95 Theses" that have sparked the Protestant Reformation, BlessU2, a robot that beams lights from its hands and grants automated "blessings" to parishioners, was developed by the Evangelical Church in Hesse and Nassau and installed in a church in the historic town of Wittenberg[9]. Sebastian von Gehren, a spokesperson from the church specified that the robot was not given a human-like appearance on purpose, so as to "inspire discussion, while not replacing the blessing of a pastor"[10]. Stephan Krebs of the Protestant church added that it was an attempt at seeing whether it is possible to be blessed by a machine, or if a human being is needed and to verify whether it was plausible to bring a theological perspective to a machine. He however quickly added that a machine "could never substitute for pastoral care".

[8] https://www.buddhistdoor.net/news/robots-take-on-monastic-roles-in-japan-and-china

[9] https://religionnews.com/2017/10/11/blessing-robots-is-a-technological-reformation-coming/

[10] www.cbn.com/cbnnews/world/2017/june/not-science-fiction-robot-pastor-will-take-the-pulpit-soon

3.2.4 Simulating Religion Project

One of the most fascinating attempts at combining a religious mindset and modern technology in the aim of predicting the future of societies is the Simulating Religion Project[11], co-run by the Center for Mind and Culture in Boston Institute for the Bio-Cultural Study of Religion, the Virginia Modeling and Simulation Center, and the Center for Modeling Social Systems at the University of Agder in Norway.

The project is based on modeling religion – turning historical theories or anthropological theories into an algorithm with various variables and an actual computer simulation. One of the goals of the project is to provide politicians an empirical tool that will help them assess competing policy options so they can choose the most effective one. One of the themes which the project focuses on is called "Modeling Religion in Norway"[12] and simulates shifts in religious and secular beliefs as agents interact with one another at different levels of education and existential security[13].

In the following section I present my first experiments performed to quantify the religious beliefs of Japanese users and automatically estimate the emotional and moral associations resulting from such beliefs.

[11] http://www.ibcsr.org/index.php/institute-research-portals/simulating-religion-project

[12] http://simrel.org/modrn/

[13] The state of Norwegian society (wealth, educational model, religious mindset etc.) is described using algorithms, which are then changed in order for instance to represent an influx of refugees from war-torn regions and the influence of this phenomenon on the rise of right-wing extremism in Norway.

3.3 Empirical Study on Presence of Buddhist Concepts in Japanese Blogs

3.3.1 State of Religious Life of Japanese

Religion in Japan is dominated by two main religions – Buddhism and Shintō. According to a survey[14] carried out in 2014, less than 40% of the population of Japan identifies with an organized religion: around 45% of those are Buddhists, 48% are members of Shintō sects and derived religions, and about 1% are Christians.

However, a characteristic trait of Japan is, in contrast to Western countries, a low level of identification with an organized religion paired with a high level of actual participation in religious rituals of both main or more than two religions – the average person typically follows the religious rituals at ceremonies like birth, wedding, and funerals, visits a shrine or temple on the New Year's day and before major life events (exams, childbirth), buys talismans and participates in local festivals (*matsuri*), most of which have a religious background. The total number of people estimated to participate in some form of religious ritual according to the Japanese Ministry of Education, Culture, Sports, Science and Technology topped 190 million people, which is more than the actual population of Japan (125,710,000 people as of July 2020).

It has not been yet measured to what degree is Buddhism present in the consciousness of the masses, or whether does what is considered ethical behavior according to Buddhist religious principles corresponds with what the general population considers to be ethically proper or improper behavior. In the following

[14] http://www.bunka.go.jp/tokei_hakusho_shuppan/tokeichosa/shumu/pdf/h26kekka.pdf

section I will analyze on a number of particular examples to what degree is religious terminology present in the vocabulary of a usual Japanese Internet user – thus what could be considered a general consciousness of the Japanese people– and what associations it generates, from the point of view of emotional life and moral implications.

3.3.2 Resources and Tools Applied in this Study

The Digital Dictionary of Buddhism[15] is the largest dictionary of Buddhist terms available online. It is a lexicon of Chinese ideograph-based terms containing of texts, names of temples, schools, persons, etc. found in Buddhist canonical sources. It also features the Chinese-Japanese-Korean-Vietnamese/English Dictionary [CJKV-E] – a compilation of Chinese ideographs, as well as ideograph-comprised compound words, text names, person names, etc., found primarily in the Confucian and Taoist classics as well as vocabulary from Neo-Confucian texts and other philosophical and historical sources. It was established in July 1995, and is updated monthly. As of December 31, 2020, the dictionary contains 74,841 entries. An example of a dictionary entry is shown in Figure 3.1.

Yet Another Corpus of Internet Sentences (YACIS [54]) is the largest Web based blog corpus available for Japanese language. It was collected automatically by Maciejewski et al. [33] from the pages of the Ameba blog service. It contains 5.6 billion words within 350 million sentences. Maciejewski et al. were able to extract only pages containing Japanese posts (pages with legal disclaimers or written in languages other than Japanese were omitted). In the initial phase they provided their crawler, optimized to crawl only the Ameba blog service, with 1000 links

[15] http://www.buddhism-dict.net/ddb/


Digital Dictionary of Buddhism

DDB Top Page DDB CJKV-E SAT Taishō XML source

佛

Pronunciations
[py] fó
[wg] fo
[hg] 붇
[mc] bul
[mr] pul
[kk] ブツ
[hi] ほとけ
[hb] butsu
[hb] hotoke
[qn] phật

Basic Meaning: buddha

Senses:

- The most basic meaning of the Sanskrit/Pāli term buddha is that of '*awakened*,' in the special sense of one who has awakened to the true nature of existence. In this sense, the meaning is often translated as '*enlightened*' 覺. One who has completely extinguished all afflictive states of mind—who utterly lacks all the mixed manifestations of ignorance 無明, desire 貪, and dislike 瞋. [Charles Muller; source(s): Ui, Nakamura, JEBD, Yokoi]
- The historical Buddha, Śākyamuni, who lived in India during the 6th century BCE, who had a major enlightenment experience, who subsequently taught others the way to achieve liberation, and following whose death, the religious movement called Buddhism, was formed. [Charles Muller]
- Any one of a number of buddhas who are said to exist in the world. In the early Pāli tradition, six are named, but in the later developing Mahāyāna tradition, buddhas are understood to be infinite in number. In this sense, anyone who attains release (*mokṣa* 解脫, *nirvāṇa* 涅槃) from this world of recurring rebirths (saṃsāra 流轉) can be called, in the appropriate contexts, a Buddha. [Charles Muller]
- Buddha as an eternal principle of enlightenment/reality, expressed in the notion of the *dharmakāya* 法身 of the Buddha. [Charles Muller]
- Buddha as Buddhahood 佛地, the final stage one attains as the result of practices. [Charles Muller]
- One of the six completions 六成就. [Charles Muller]
- Buddha, from *budh* to 'be aware of,' 'conceive,' 'observe,' 'wake' ; also written 佛陀; 浮圖; 浮陀;浮頭; 浮塔; 勃陀; 勃馱; 沒馱;母馱; 母陀; 部陀; 休屠. Buddha means 'completely conscious, enlightened,' and came to mean the enlightener. The Chinese translation is 覺 to perceive, be aware, awaken; and 智 gnosis, knowledge. There is an Eternal Buddha, see e.g. the *Lotus Sutra*, cap. 16, and multitudes of buddhas, but the personality of a Supreme Buddha, an Ādi-Buddha, is not defined. Buddha is in and through all things, and some schools are definitely

Figure 3.1: Example of a dictionary entry from the Digital Dictionary of Buddhism

taken from Google (response to one simple query: 'site:ameblo.jp'). They saved all the pages to a disk as raw HTML files (each page in a separate file) and afterward extracted all the posts and comments and divided them into sentences. The original structure (blog post and comments) was preserved, thanks to which semantic relations between posts and comments were retained. The blog service from which the corpus was extracted (Ameba) is encoded by default in Unicode, thus there was no problem with character encoding. It also has a clear and stable HTML meta-structure, thanks to which they managed to extract metadata such as the blog title and author. The corpus was first presented as an unannotated corpus. Later, Ptaszynski et al. annotated it with syntactic information, such as POS, dependency structure or named entity tags. An example of the original blog structure in XML is represented in Figure 3.2. Some statistics about the corpus are represented in Table 3.1.

# of web pages	12,938,606
# of unique bloggers	60,658
average # of pages/blogger	213.3
# of pages with comments	6,421,577
# of comments	50,560,024
average # of comment/page	7.873
# of words	5,600,597,095
# of all sentences	354,288,529
# of words per sentence (average)	15
# of characters per sentence (average)	77

Table 3.1: Statistics of YACIS

```
<doc url="http://ameblo.jp/capo-del-rosso/entry-000000.html" time="2009-12-05 21:11:46" id="2000001">
        <post>
                <s>今日から十月です。</s>
                [Its October from today.]
                <s>なんか、九月はいつもよりアッという間に過ぎたような気がするなぁ。</s>
                [I have a strange feeling September passed faster than usual.]
                ...
        </post>
        <comments>
                <cmt>
                        <s>色々と忙しいですね～！</s>
                        [Oh, you've been busy, weren't you?]
                        ...
                </cmt>
                <cmt>
                        <s>お疲れｻﾏです(^O^)</s>
                        [Well done! Cheers for good work (^o^)]
                        ...
                </cmt>
        </comments>
</doc>
```

Figure 3.2: An example of the original blog structure in YACIS in XML (reproduced from [54])

The reasons for choosing YACIS as my corpus of choice were the following. In quantitative studies it is very important to provide a statistical reliable random sample of sentences or documents as a dataset for analysis. The larger is the source, the more statistically reliable is the dataset. Since YACIS contains 354 million sentences in 13 million documents, it can be considered sufficiently reliable for the task of dataset extraction for various quantitative studies, as probability of extracting twice the same sentence is close to zero. Moreover, as mentioned in the Introduction, YACIS remains the largest single domain corpus for Japanese language.

The Automatic Moral Judgement Agent Based on Wisdom of WebCrowd and Emotions (hereafter Moral Reasoning Agent or MRA), was first proposed by Rzepka and Araki [59] in 2005 and further developed by Komuda et al. [28]. The moral consequence retrieval agent was based on the idea of Wisdom of Crowd. In particular Komuda et al. used a Web-mining technique to gather consequences of actions applying causality relations, to extract from the Web emotional and ethical consequences of actions found in input.

The agent takes a sentence as an input and in a specified corpus (such as the Internet or blog corpus as above) searches for emotion types and morality-related concepts associating with the sentence contents. The technique is composed of five steps: a) extraction of input phrase; b) modification of the phrase with causality morphemes; c) searching for the modified phrase in the specified corpus; d) matching to the predetermined lexicons[16] (containing moral and emotional concepts) and

[16]The emotive lexicon was based on Nakamura's Dictionary of Emotive Expressions ([43]) and ML-Ask ([53]). It contains 2100 items (words and phrases) describing emotional states, divided into 10 emotions classes. The ethical consequence lexicon was created through substituting the ten emotion types into five pairs of word groups representing Kohlberg's stages of moral development (based on Lawrence Kohlberg's theory of human moral development, in which he assumed successive changes in aspects by which I consider an action good or wrong). The items in

extraction of emotion associations; e) ranking creation and output.

A model of the functioning of the Moral Reasoning Agent is demonstrated in Figure 3.3.

The modified phrases are queried in the corpus (e.g. the Internet) for a set specified number of snippets for one modified phrase (default is 300 per phrase). This way a large number of snippets for each queried phrase is extracted from the Web and cross-referenced with the emotive and moral lexicons. For example, a phrase "thank you" is likely to associate with gratitude, relief and joy and "saving a person" is more likely to associate with such moral associations as "deserve praise", or "thankfulness" and ethical consequences such as "be praised by people" while "killing a person" is more likely to be linked with such ethical consequences as "going to jail", or "condemn".

The agent was tested on over 100 ethically significant real-world problems, such as "killing a man", "stealing money", "bribing someone", "helping people" or "saving environment". The problems in a form of sentences, or statements, were first annotated with probable moral consequences by laypeople. When compared to these annotations, the agent's results were correct in approximately 86% (accuracy). Some examples of the results are presented in Figure 3.4.

In our research I applied the agent to search for emotional and moral associations correlating with most commonly appearing phrases containing Buddhist terminology.

---

the lexicon were distributed as follows: Praises (18) / Reprimands (33); Awards (25) / Penalties (15); Society approval (8) / Society disapproval (8); Legal (8) / Illegal (8); Forgivable (6) / Unforgivable.

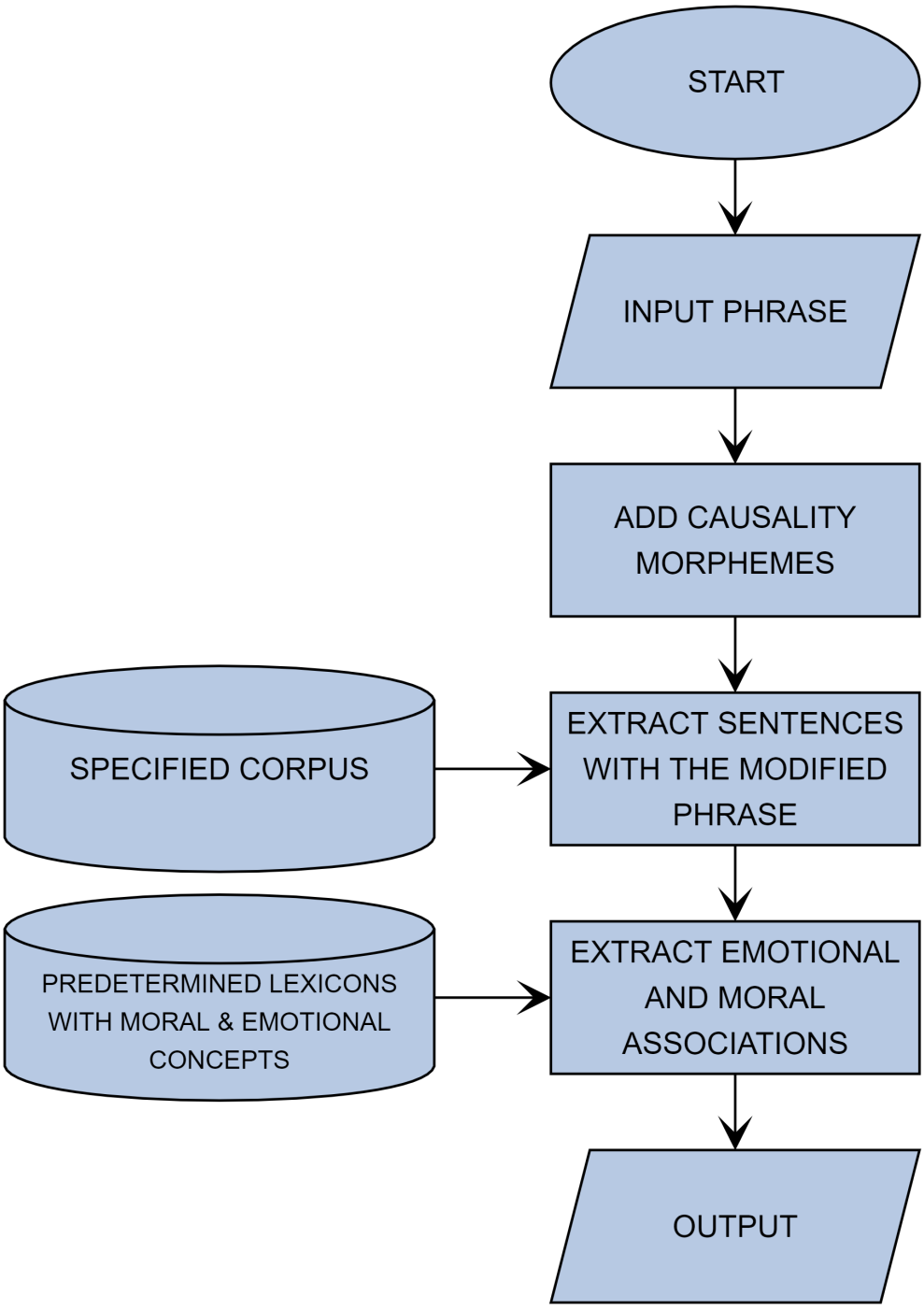

Figure 3.3: A model of the functioning of the Moral Reasoning Agent

emotional conseq.	**results**	**score**	**ethical conseq.**	**results**	**score**
"To hurt somebody."					
anger	13.01/54.1	0.24	penalty/	4.01/7.1	0.565
fear	12.01/54.1	0.22	punishment		
sadness	11.01/54.1	0.2			
"To kill one's own mother."					
sadness	9.01/35.1	0.26	penalty/	5.01/5.1	0.982
surprise	6.01/35.1	0.17	punishment		
anger	5.01/35.1	0.14			
"To steal an apple."					
surprise	2.01/6.1	0.33	reprimand/	3.01/3.1	0.971
anger	2.01/6.1	0.33	scold		
"To steal money."					
anger	3.01/9.1	0.33	penalty/punish.	3.01/6.1	0.493
sadness	2.01/9.1	0.22	reprimand/sco.	2.01/6.1	0.330
"To kill an animal."					
dislike	7.01/23.1	0.3	penalty/	36.01/45.1	0.798
sadness	5.01/23.1	0.22	punishment		

Figure 3.4: Example of the results for queries presented to the Moral Reasoning Agent (reproduced from [28])

3.3.3 Initial Phrase Extraction

In the analysis procedure I first checked how many of the headwords from the Digital Dictionary of Buddhism appear in the YACIS corpus. Among headwords from the dictionary, 22,397 Buddhist terms or terms associated with Buddhism appeared at least once in YACIS. From these terms, the most popular one appeared in the corpus a total of 27,884 times (麻布 [jap. *azabu*, "linen"]), while on the other hand 3,616 least frequent terms appeared only once. In total a number of 15,087,745 sentences from various blog posts (about 4% of all sentences) contained Buddhism-related terms.

However, it must be noted that among Buddhist-related terms appearing in YACIS a majority were not strictly religious terms, but rather terms related to the Buddhist religion to some extent, like the term with the largest number of hits, which are part of regular Japanese vocabulary and also appear in Buddhist

scriptures, and thus are considered related to Buddhism by the Digital Dictionary of Buddhism. Ten words with the highest number of hits are shown in Table 3.2.

Term	Reading	Meaning	Occurrences in YACIS
麻布	azabu	linen	27,884
時起	Jiki	monk name	25,391
株	shu	stump	24,404
胡麻	goma	sesame	23,713
日日	nichi-nichi	everyday	23,553
白	shiro	white	23,170
入門	nyūmon	to become a disciple	23,109
起動	kidō	move, awakening	23,015
右	migi	right (direction)	22,808
海水	kaisui	sea water	22,635

Table 3.2: Ten Buddhism-related words with the highest number of hits in YACIS

3.3.4 Filtering of Phrases Used in Strictly Buddhist Context

After having specified which Buddhist terms appear most often in general, for further detailed analysis I needed to specify the seed phrases, which typically appear in blogs in a Buddhist context, and not as lexicalized expressions. These phrases would be further applied to the Moral Reasoning Agent to find surrounding emotional and moral associations. I selected the words covering 83.5% of all results, which only amounted to 7.5% of the total vocabulary found in blogs (see Figure 3.5). This means that only the top 1,727 words generated an output of 12,070,196 posts (83.5% of the total number of posts, from the overall 15,087,745 posts).

I noticed that much of the vocabulary – especially terms composed of one and two characters – covered common words, such as "egg" (卵 [*tamago*], 17,842 hits), which appeared in the Buddhist canon but carried no specifically Buddhist

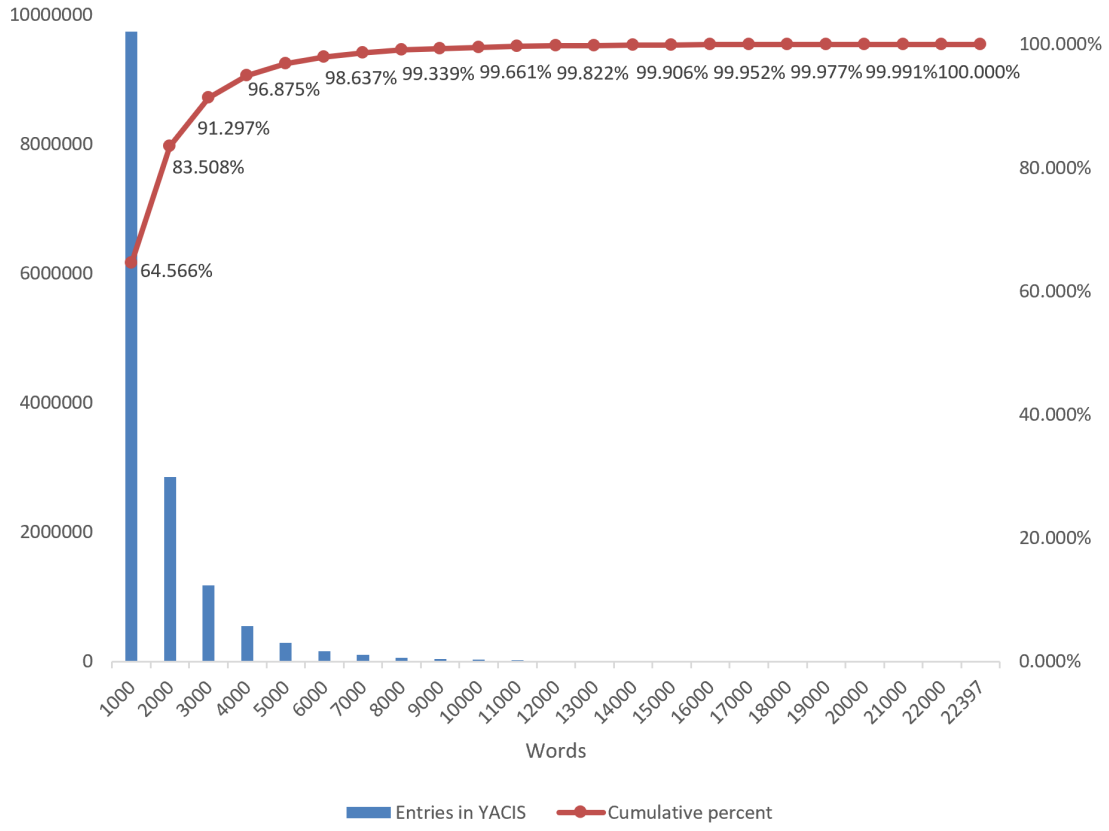

Figure 3.5: Top 1727 Buddhism-related words in YACIS covering 83.5% of the results

meanings in everyday life.

In order to only leave out words entries with a strictly Buddhist meaning, I performed a cross check of headwords from the Digital Dictionary of Buddhism with Wikipedia (using all the 912 headwords[17] in Japanese belonging to the category "Buddhism" in Wikipedia). This left us with 215 terms that belonged both to the category of religious Buddhist vocabulary and were present in Wikipedia.

Out of the most popular 1,727 terms, a further cross-check with Wikipedia proved that only 25 terms appeared in all three resources, namely the Digital Dictionary of Buddhism, YACIS and the Japanese Wikipedia. Among those, 22 were composed of two characters, while there were two terms composed of one character (有 [*yū*, "presence, existence"] and (華 [*hana*, "flower"]) and one term composed of three characters:四天王(*shitennō*, "Four Heavenly Kings", the four guardian gods, who protect the four quarters of the universe). The terms composed of two Chinese characters covered, among others, basic concepts in Buddhist philosophy (外道 [*gedō*, "blasphemer"], 輪廻 [*rinne*, "transmigration"] or 他力 [*tariki*, "the outer power" – the attainment of liberation in reliance on the salvific powers of a great buddha or bodhisattva]), ranks in monkhood (法師 [*hōshi*, a teacher of the Dharma], 僧侶 [*sōryo*, Buddhist monk], 上人 [*shōnin*, a monk of superior wisdom, virtue, and conduct]), Buddhist rituals (供養 [*kuyō*, "offering"], 布施 [*fuse*, "almsgiving"], 法要 [*hōyō*, a funeral ceremony]), elements indispensable in Buddhist liturgy (袈裟 [*kesa*, "vestment" that is worn draped over the left shoulder by Buddhist monks in East Asia]) or terms belonging to the category of temple organization and Buddhist architecture: 寺院 [*ji'in*, "Buddhist temple"] or 本山 [*honzan*, head temple or main temple].

[17]State of Wikipedia contents as of September 10, 2018

The results in the case of two terms, however Buddhist, proved irrelevant for this research due to their usage in contemporary Japanese language – the term 六時 [*rokuji*] was originally meant to describe six periods of the day, devoted to different activities in a monastery. However in modern-day Japanese it simply means "six o'clock" and all found results related to the modern use of the word. Same goes for the term 五輪 [*gorin*, literally "five rings"] originally meaning "five members of the body", "five foundations of the world" or "five fingers of the Buddha", but is used in Japanese nowadays to designate the Olympic Games.

Surprisingly enough, basic and most obvious Buddhist terms such as *hotoke* (仏, "Buddha"), *bosatsu* (菩薩, "bodhisattva", a Buddhist practitioner intent on the attainment of enlightenment based on profoundly altruistic motivations) or *gō* (業, deeds and their effects on the character, the law of karma) did not make it into the final list.

3.3.5 Filtering of Word n-grams for Emotional and Moral Association Extraction

To be able to find emotional and moral associations for the selected Buddhist terms, I needed to specify short meaningful phrases (three to four word-long) in which such terms usually appear, to be further used as input in the Moral Reasoning Agent (MRA). Since the data still accounted for a large portion of text, I decided to specify such phrases semi-automatically. The procedure for selecting those phrases was as follows: 1) I used all sentences in which the applied terms appeared. 2) I tokenized the sentences (divided into words) for further automatic extraction of frequent n-word-long phrases (n-grams, in this research limited to 3-grams and 4-grams), and 3) lemmatized them (generalized their grammatical

forms to generic dictionary forms) for more robust search. Then, 4) I extracted all 3-grams and 4-grams from the sentences containing the phrases, and 5) retained only those n-grams that actually contained the searched word. Finally, 6) I applied an additional filter to retain only those phrases that contained not only the word in question but also grammatical particles, which are used by MRA in its Web-mining procedure. This additional filter gave me the certitude that the phrases would certainly appear in a meaningful sentence, which would make the extraction of emotional and moral concepts more feasible.

3.3.5.1 Tokenization and Lemmatization

After extracting the entries from YACIS containing the most popular Buddhist vocabulary composed of two, three and four characters, I used the MeCab, a standard morphological analyzer for Japanese[18]. An example of tokenization and lemmatization was shown in Figure 3.3.

3.3.5.2 Filtering of N-grams Containing Grammatical Particles

Next, I proceeded to extract word n-grams, leaving only those that contained both the researched terms and grammatical particles belonging to the categories of case markers (格助詞 [*kaku-joshi*]), namelyが [*ga*], の [*no*], を [*wo*], に [*ni*],へ [*he*], と [*to*], で [*de*], から [*kara*] and より [*yori*], adverbial particles (副助詞 [*fuku-joshi*]): ばかり [*bakari*], まで [*made*], だけ [*dake*], ほど [*hodo*], くらい [*kurai*], など [*nado*], なり [*nari*], やら [*yara*], binding particles 係助詞([*kakari-joshi*]): は [*wa*], も [*mo*], こそ [*koso*], でも[*demo*], しか [*shika*], さえ[*sae*], だに [*dani*] and conjunctive particles (接続助詞 [*setsuzoku-joshi*]): ば [*ba*], や [*ya*], が [*ga*], て [*te*], のに

[18]http://taku910.github.io/mecab/

Original entry from YACIS
自分にたくさん嘘をつくだけならまだしも、色んな人に嘘をついた気がする。それも全部自分のための嘘だったと思う。環境が変わったから、っていうのもあるかもしれないけど...必要以上の嘘をついた。来年は、嘘をつかないではっきりと言えたらいいと思う。思う、じゃあなくて、そうであればいい。そうでなくてはいけない。

Translation
It would be lesser evil if I would just keep on lying to myself, but I feel like I have been lying to many people. Moreover,I think it was all lies in my own interest. Well, you could probably say it is also because my environment has changed... but I lied more than it was necessary. Next year, I think it would be good if I say clearly that I will not lie anymore. No, more than thinking, it should be this way. It cannot be another way.

After tokenization and lemmatization
自分 に たくさん 嘘 を つく だけ だ まだしも 、 色んな 人 に 嘘 を つく た 気 が する 。 それ も 全部 自分 の ため の 嘘 だ た と 思う 。 環境 が 変わる た から 、 っていう の も ある かも しれる ない けど ... 必要 以上 の 嘘 を つく た 。 来年 は 、 嘘 を つく ない で はっきり と 言える た いい と 思う 。 思う 、 じゃ あ ない て 、 そう だ ある ば いい 。 そう だ ない て は いける ない 。

Table 3.3: An example of tokenization and lemmatization of an entry from YACIS performed using MeCab

[*noni*], ので [*node*], から [*kara*], ところが [*tokoroga*], けれども [*keredomo*]. The total number of n-grams left to consider upon performing this preliminary cleaning was 181,785 3- and 4-grams. Next, I proceeded with leaving out only the top 10 n-grams with the highest number of hits for the 25 terms appearing in all three ressources used in this experiment and removing any superfluous spaces and numbers. This left us with 332 results to consider (overall number of n-grams for all 25 terms to be used as input in MRA).

3.3.6 Examining Moral and Emotional Associations of Buddhist Terms

With the remaining 3- and 4-grams containing the Buddhist term in question and a particle, obtained in the previous step, I checked representative examples of expressions containing Buddhist vocabulary using the Moral Reasoning Agent. In the future I plan on making the process of checking all n-grams automatic and extracting an average value. However, this time as a proof of concept and for the sake of further discussion I only chose a few terms denominating basic Buddhist concepts and the eventual discrepancies between how they are understood and explained in Buddhist philosophy and how they are perceived by general population (in this case by Japanese blog users).

3.3.6.1 *Rinne* or Transmigration

One of the cornerstone concepts associated with the Buddhist religion is transmigration, *rinne* (輪廻), the belief that all living things repeatedly pass through life and death, like a continually spinning wheel (falsely known as reincarnation). In Buddhism, life is inevitably connected with suffering and such being the case, you

want to escape the process of transmigration by attaining enlightenment. However, Japanese blog users seem to be quite resigned to their fate, with 129 results of moral associations with the term equally divided between good (with an example phrase stating: 輪廻転生は「忍耐」を与えられる事それに立ち向かい、その先に幸せがある – "When it comes to transmigration and rebirth, it gives you perseverance to face things, and at the end of this process happiness awaits") and bad (本当は輪廻転生の生と死の繰り返しから抜け出せないことは良くないことと考えられている – "In fact, I believe that it is not a good thing not being able to escape the cycle of death and rebirth and transmigration"), another 16% of the results indicating transmigration was a phenomenon that was "the right thing to happen" and a further 5% even stating it was "worth encouraging". However, there were also 29 emotional consequences mostly described as "bad" or "negative", with phrases including emotive terms such as "suffering" or "suicide", giving an insight into the distress that transmigration might bring an individual.

	Term	輪廻転生 [*rinne tensei*, "transmigration"]
	Number of n-grams	1,287 3-grams and 4,522 4-grams
	Sample n-gram	輪廻転生の [*rinne tensei no*, "transmigration" + possessive particle]
	Frequency in the corpus	129 times
	Social consequences	positive: 78%, negative: 21%
	Emotional consequences	positive: 44%, negative: 55%
Moral categories	Correct action	16% of the results
	Worth encouraging	5% of the results
	Worth praise	n/a
Immoral categories	Illegal act	0% of the results
	Unacceptable conduct	5% of the results
	Worth reprimanding	5% of the results

Table 3.4: Results of the analysis for the term *rinne tensei*

3.3.6.2 *Gedō* or Blasphemer

Results of analyzing the term 外道 (*gedō*, "blasphemer"), a follower of a religious/philosophical tradition that does not accept basic Buddhist positions, surprisingly enough demonstrated that among the 66 examples of the usage of the term 外道衆, (*gedōshū*, "blasphemous crowd" – although the term was often used metaphorically, signifying a rebellious group)- 100% of the results indicated it was a socially positive phenomenon, while 64% of the results also stipulated that belonging to such a crowd was an emotionally positive act. 18% thought that "being part of the blasphemous crowd" was a "morally proper act", while simultaneously 0% found it to be something "in accordance with the law" or "worth praise". An analysis of the emotional response to this term seemed to indicate that with 14 positive emotional associations "most people find pleasure (in being part of the blasphemous crowd – or in distancing themselves from the blasphemous crowd)". I also came across numerous examples of the usage of the term *gedō* as it is now understood in contemporary Japanese: "a way of doing things which is not the canonical way", as exemplified by sentences such as: 昔、バス釣りの外道で釣ったミドリガメ ("This is a green turtle that I caught a long time ago while illegally fishing for bass"). Admirers of certain genres of animated films, especially, were prone to calling any critics of their pastime of choice "blasphemers" (こんな考えをする私は外道衆の仲間入りですかね – "Does thinking in this way makes me one of the blasphemers?").

	Term	外道衆 [*gedōshū*, "blasphemous crowd"]
	Number of n-grams	1,841 3-grams and 5,487 4-grams
	Sample n-gram	外道衆が [*gedōshū ga*, "blasphemous crowd" + nominative particle]
	Frequency in the corpus	66 times
	Social consequences	positive: 100%, negative: 0%
	Emotional consequences	positive: 64%, negative: 35%
Moral categories	Correct action	18% of the results
	Worth encouraging	n/a
	Worth praise	n/a
Immoral categories	Illegal act	n/a
	Unacceptable conduct	n/a
	Worth reprimanding	n/a

Table 3.5: Results of the analysis for the term *gedōshū*

3.3.6.3 *Tariki* or Outer Force

Tariki (他力, "other/outer force") a term crucial for Amidistic Buddhist schools in Japan[19], implying that the only attainable salvation from the suffering of this realm comes through belief in the grace of Buddha Amitâbha[20] (as opposed to *jiriki*, 自力, "one's own spiritual power"), was present in 29 examples. The social consequences were here again equally divided between "good" and "bad" – mostly in sentences such as 自力では敗退ですが、他力での復活の可能性ありますね – "On your own you stand no chance, but with outside help there's a chance for revival". An equal 25% of the results indicated it was a "correct" or "improper" approach to life.

[19]The Jōdo-shū ("Pure Land school"), founded by Hōnen in 1175 and the the Jōdo-shinshū ("True Pure Land school") founded by Shinran taught relatively simple methods of recitation of the Buddha Amida's name for the purpose of attaining rebirth in the Western Heaven (Pure Land) and together became the most popular form of Buddhism among the common people and lay practitioners in Japan

[20]The Buddha of Limitless Light or Limitless Life. A Buddha who possesses infinite meritorious qualities and who expounds the Dharma in his pure paradise (Sukhavati) in the West. Amitâbha is the primary deity of the Pure Land schools of Buddhism which developed and spread in China, Vietnam, Korea, and Japan.

	Term	他力 [*tariki*, "other/outer force"]
	Number of n-grams	847 3-grams and 2,906 4-grams
	Sample n-gram	他力で [*tariki de*, "through outer force"]
	Frequency in the corpus	29 times
	Social consequences	positive: 50%, negative: 50%
	Emotional consequences	positive: 50%, negative: 50%
Moral categories	Correct action	25% of the results
	Worth encouraging	n/a
	Worth praise	n/a
Immoral categories	Illegal act	n/a
	Unacceptable conduct	25% of the results
	Worth reprimanding	n/a

Table 3.6: Results of the analysis for the term *tariki de*

3.3.6.4 *Fuse* or Almsgiving

Japanese blog users also revealed the belief in negative ethical implications of the custom of *fuse*, 布施, almsgiving, in present day Japan most often in the form of money offerings to Buddhist temples. Among the 40 examples of the usage of the term *fuse* found in Ameba blogs, the total of social consequences was given as 100% negative (e.g. 実際に徳を積むとは無償の行為、労働、お布施をすること金品を奉納することになっています- In fact accumulating virtue ought to be an action for which you don't expect compensation and when you work or give alms, you actually obtain financial gains), and 55% of the results indicated that almsgiving was an unethical action.

3.3.6.5 *Kuyō* or Veneration

A similar Buddhist term is 供養 (*kuyō*), signifying "veneration", an offering of food, drink, clothing etc. to a buddha, monk or teacher, or a special commemorative service held to mark such things as the construction of a temple or donations made to individual monks. I have found 179 examples of the usage of the term in Ameba. Here however, the total of emotional consequences (84 kinds of emotions were

	Term	布施 [*fuse*, "almsgiving"]
	Number of n-grams	1,199 3-grams and 3,856 4-grams
	Sample n-gram	布施をする [*fuse wo suru*, "perform almsgiving"]
	Frequency in the corpus	40 times
	Social consequences	positive: 0%, negative: 100%
	Emotional consequences	positive: 50%, negative: 50%
Moral categories	Correct action	n/a
	Worth encouraging	n/a
	Worth praise	n/a
Immoral categories	Illegal act	20% of the results
	Unacceptable conduct	n/a
	Worth reprimanding	n/a

Table 3.7: Results of the analysis for the term *fuse*

associated with the term) was good (61% of the results). 62% of the results found it to be an ethical thing to do, with a further 7% of the results indicating it was "proper", with proof found in sentences such as : 元気を出してくださいね。それが一番の供養だと思います。 – "Please try to feel better, this is the best offering you can give them [deceased members of the family]".

	Term	供養 [*kuyō*, "offering" or "commemorative service"]
	Number of n-grams	3,365 3-grams and 13,014 4-grams
	Sample n-gram	への供養 [*he no kuyō*, "commemorative service for [someone]"]
	Frequency in the corpus	179 times
	Social consequences	positive: 71%, negative: 28%
	Emotional consequences	positive: 61%, negative: 38%
Moral categories	Correct action	7% of the results
	Worth encouraging	1% of the results
	Worth praise	n/a
Immoral categories	Illegal act	n/a
	Unacceptable conduct	2% of the results
	Worth reprimanding	2% of the results

Table 3.8: Results of the analysis for the term *kuyō*

3.3.6.6 Attitudes Toward Becoming a Monk – *Soryō*

Buddhism in its beginnings was a monastic religion and in many predominantly Buddhist countries up to this day young men spend some time practicing as Buddhist novice monks, as it is thought to be a way of repenting for the sins of one's ancestors and accumulate good deeds. However, when it comes to opinions and judgments of the Japanese blog users about Buddhist clergy, responses seem more complex. I have found 45 examples of the expression 僧侶になる (*soryō ni naru*, "become a monk") in Ameba. The total of social consequences was good at a 100%, indicating that becoming a monk was socially encouraged; however, the total of emotional consequences (11 examples) was bad at 72% – as a proof I can mention sentences such as: 僧侶になるのも大変だけど、なってからも大変なんだな – "It's an arduous process to become a monk and it's also hard once you became one". Furthermore, 66% of the responses implied it was not an ethically proper action. Examples of sentences included phrases such as: しかし、この僧侶の方はソレ＝「煩悩」だと感じてしまったわけですよね ("However I could feel that this monk was a sinner").

	Term	僧侶 [*sōryo*, "monk"]
	Number of n-grams	3,674 3-grams and 11,469 4-grams
	Sample n-gram	僧侶になる [*sōryo ni naru*, "become a monk"]
	Frequency in the corpus	45 times
	Social consequences	positive: 100%, negative: 0%
	Emotional consequences	positive: 27%, negative: 72%
Moral categories	Correct action	n/a
	Worth encouraging	25% of the results
	Worth praise	n/a
Immoral categories	Illegal act	n/a
	Unacceptable conduct	n/a
	Worth reprimanding	n/a

Table 3.9: Results of the analysis for the term *sōryo*

3.4 Discussion

In the outcome of this experiment, further filtering of the obtained results proved to be a key improvement in order to obtain a more reliable output. The sample obtained upon cross-checking the terms collected through initial analysis with headwords from Japanese Wikipedia in the "Buddhism" category was small, but contained terms belonging to key categories of Buddhist vocabulary, such as philosophical rudiments or monastic organization.

Several terms have had a religious meaning originally, which was lost or forgotten over time, although those terms still appear in a dictionary of Buddhist terminology. In the future I plan to verify which dictionaries of Buddhist terminology make a distinction between terms still widely in use and those that have shifted meanings or became obsolete. This would solve one of the problems I encountered with the obtained sample, that proved irrelevant for the research as their Buddhist origin has become obscure. This regards specifically two results, namely, the term 六時 (*rokuji*), originally meant to describe six periods of the day devoted to different activities in a monastery, that in modern Japanese came to mean "six o'clock", and the term 五輪 (*gorin*, literally "five rings") originally meaning "five members of the body", "five foundations of the world" or "five fingers of the Buddha", but is used in Japanese nowadays to designate the Olympic Games.

In the future I would like to deepen my research through corroborating the list of vocabulary obtained by filtering it through a list of Buddhist terms used in everyday modern Japanese. This would make it possible firstly to check, how many terms are nowadays used regardless of their religious provenience and secondly to evaluate what percentage of vocabulary and in which categories (such as proverbs, art-related terms, etc.) has Buddhist origin. A bigger database of Buddhist terms

as used in Japanese in the past and nowadays would also make it possible to track changes occurring in the structure of the Japanese vocabulary.

A recent survey done by the consulting services corporation Deloitte[21] proves that consumers and users of machines have become at the same time the critics and the creators and expect to have an influence on the shape of the products and services they use. For a religious user of a given service or intelligent machine, its level of personalization could be the key to assuring the user's satisfaction, creating an engaging relationship between the product and the user.

One might think that given the dislike of the Japanese for identifying themselves with an organized religion, the Japanese society is profoundly secular. However, our study has demonstrated that Buddhism as a topic is present in modern, secular media channels such as blogs. As such, there arises the need among designers of future technologies in Japan (but also anywhere else worldwide) to take into account plausible religious preferences and opinions of its users. However, our study also proves that what is defined as a canon of moral or ethical principles for followers of religions such as Buddhism does not necessarily overlap with what the larger population (here: Japanese blog users) would describe as their Buddhism-backed morality or ethical compass. Thus, there arises the need of regularly updating the state of knowledge on the religious outlook of the modern society to make future Intelligent Machines capable of taking into account user's religious worldview. Such personalized machines need to be able to trace the dynamic changes that occur on the crossroads of religion and the post-industrial civilization.

[21] https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/consumer-business/deloitte-uk-consumer-review-mass-personalisation.pdf

3.5 Conclusions

In this chapter I presented a preliminary analysis on whether and to what degree religious vocabulary is present in the everyday of Japanese Internet users (in this case in a repository of blog entries) and thus in the general conscience of people. I was specifically interested in discovering what is the emotional response of bloggers to Buddhist terminology and what are the results of ethical evaluations within snippets of texts including Buddhist vocabulary. In order to do this, I checked which headwords from the Digital Dictionary of Buddhism (the largest lexicon of Buddhist terms available online) appeared in YACIS, a large-scale corpus of Japanese language based on blog entries from the Japanese Ameba blog hosting service.

After obtaining the first data, I focused on the most popular terms, covering 83.5% of all results (5% of the total vocabulary and an output of 12,070,196 posts). I then noticed that although a lot of headwords from the Digital Dictionary of Buddhism appeared in YACIS, the majority of those were in fact expressions that are part of regular Japanese vocabulary, and although appear in Buddhist scriptures, are not specifically used in religious context. To solve this problem, I then further filtered the phrases in order to exclusively retain vocabulary with a strict Buddhist meaning.

Next, I tokenized and lemmatized the sentences containing Buddhist vocabulary, extracted n-grams (up to 4-grams containing the term in question and a grammatical particle) and finally checked a few most interesting and representative examples of expressions containing Buddhist vocabulary, using an Automatic Moral Judgement Agent Based on Wisdom of WebCrowd and Emotions (a moral consequence retrieval agent that was based on the idea of Wisdom of Crowd, using a Web-mining technique

to gather consequences of actions applying causality relations, to extract from the Web emotional and ethical consequences of actions found in input).

Analysis of the data with the Moral Reasoning Agent revealed that Buddhist terms were in fact not absent as a theme from Japanese blogs and generated a strong emotional response. I chose to discuss in-depth a few terms denominating basic Buddhist concepts and the discrepancies between how they are understood and explained in Buddhist philosophy and how they are perceived by Japanese blog users.

Chapter 4

Does Change in Ethical Education Influence Core Moral Values? Application in Automatic Moral Reasoning

4.1 Introduction

While looking for roots to the formation of people's ethical outlook, I noticed that It has not been yet measured to what degree the change in the way children are introduced to moral principles through the school curriculum is reflected in the broader mindset of the population and visible in associations with certain ethical categories and societal key concepts before and after major historical events, such as for instance the Second World War.

In this chapter, I aimed at verifying how much ethics, as taught in Japan in

the course of the mandatory school curriculum before the Second World War and as taught right now, has in common with the moral attitudes of contemporary Japanese and the way common people approach core moral questions (such as those concerning the sacredness of human life). I was also interested in finding out whether a big shift in the contents of lessons taught as part of the subject "Ethics" in schools would resonate within society by influencing what the larger masses think is morally acceptable or not.

I tried to analyse this by selecting textbooks used to teach ethics at Japanese schools between 1937 and 1939, and those used in Japanese junior high schools today (2021) and analyzed what emotional and moral associations were generated by the contents of the textbooks. The analysis was performed with an automatic moral and emotional reasoning agent and based on the largest available text corpus for the Japanese language as well as on the resources of a Japanese digital library.

The remainder of this chapter is structured as follows. I give a brief outline of Japanese educational reforms and their influence on teaching ethics in Section 2, introduce the resources used in this study in Section 3, and describe the method used to analyze obtained data in Section 4, followed by the results of the experiment. Section 5 consists of a discussion and proposed directions for further research. Finally, the paper is concluded in Section 6.

4.2 Hundred Years of Japan's Educational System

Japan has a long history of compulsory schooling starting as early as 1886, fourteen years after the issuance of the 1872 Education Code, which was the

globally first comprehensive plan for mass schooling. Elementary School Regulations (*Shōgaku Kyōsoku*, 小学教則), issued in September 1873, became the impulse to introduce a new school subject called *shūshin* (修身, "teaching of good virtue").

In 1890, the Imperial Rescript on Education (*Kyōiku ni Kansuru Chokugo*,教育に関する勅語) was signed to articulate government policy on the guiding principles of education in the Empire of Japan. The main objective of the Rescript was the education of an "ideal imperial subject"[1]- it promoted core moral values such as piety, loyalty, friendship, benevolence, sincerity, building of prosperity, respect, courage, modesty, obedience, docility, conformity and submission of children to parents and teachers, nation, and the emperor.

There were six original subjects taught in the 1880s: *shūshin*, reading, writing, Japanese calligraphy, mathematics and physical education, among which *shūshin* was considered the most important subject. As of 1910, more than 98% of primary school-aged children attended a four year compulsory primary school.

After Japan's defeat in the Second World War, the "Order of Suspension of Courses in Morals, Japanese History and Geography" by the American-led General Headquarters terminated *shūshin* in December 1945 as part of a transition from pre-war militarist policies to a new law. In 1947, following the end of the Second World War, a new law, the Fundamental Law of Education (FLE), the central philosophy of which was to confirm equal rights to education and to democratize the country, was brought into existence. In terms of subjects taught at school, pre-war worship of the Emperor taught through *shūshin* was replaced with an alternative subject, *dōtoku* (道徳, roughly translatable as "ethics"), with promoted core moral

[1]The 315-character document was read aloud at all important school events, and students were required to study and memorize the text. A translation of the edict can be obtained from: https://en.wikisource.org/wiki/Imperial_Rescript_on_Education

values including self-sacrifice for the good of others, conformity over individuality (derived from Confucianism) as well as self-awareness and the development of moral thinking [16].

"The General Course of Study" in 1951 showed the direction of moral education not as a subject taught in school but as a subject interwoven through the whole school education.

However, the "Report of the Curriculum Council: Establishment of Special Time of Moral Education," in 1958, pointed out that the whole-school approach did not produce effective education and suggested that a period for moral education should be established formally.

The present Basic Act on Education[2] mentions that education should "cultivate morality and ethics". MEXT[3] started to provide free supplemental learning materials for teaching Ethics called *Kokoro no Nōto* (心のノート)or "Notebook for the Heart", in 2002.

In 2006, the Fundamental Law of Education was revised for the first time in 59 years. A major controversy resulted from its emphasis on "tradition," "discipline," "morality," and most of all, "patriotism" or "love of nation," all of which seem reminiscent of the 1890's Imperial Rescript on Education.

Further change in the way ethics are taught in Japanese schools is ahead as the Central Council for Education, an advisory body for the education minister, has submitted a report calling for upgrading moral education to an official school subject [60]. Beginning in fiscal 2018 and 2019, it was introduced on a par with traditional subjects like Japanese, mathematics, social studies and science at elementary schools

[2]A translation of the Basic Act on Education can be found here:https://www.mext.go.jp/en/policy/education/lawandplan/title01/detail01/1373798.html

[3]The Japanese Ministry of Education, Culture, Sports, Science and Technology

and junior high schools.

In this chapter, I will analyse ethics as taught in Japan before WWII and today to verify how much the pre-WWII moral attitudes have in common with those of contemporary Japanese, to what degree what is taught as ethics in schools today laps over the general population's understanding of ethics and finally to try to assess whether a major reform of the guidelines for teaching the school subject "Ethics" at school after 1946 has changed the way common people approach core moral questions – and what part of ethics, if any, was left missing after the education system reforms.

4.3 Applied Resources

4.3.1 Language resources to analyse

4.3.1.1 *Shūshin kyōjuroku*

Shūshin kyōjuroku (修身教授録, "Records from lectures in moral education") [41] is a record of the content of the seventy nine classes that Mori Nobuzō, a Japanese philosopher and educator, gave on the subject of *shūshin* between 1937 and 1939 at what is now the Osaka University of Education. Although the participants of the course were to become teachers, he conducted a course on general ethical matters, such as "the question of life and death" or "one only lives once," etc., amply enlarged through incorporating stories based on the life of role models for future educators. The book is a long seller sold in over 150 000 copies, was first published in 1989 and had over 50 new editions.

4.3.1.2 *Watashitachi no dōtoku*

Watashitachi no dōtoku (私たちの道徳, "Our morals") [37] is a textbook for the subject "Ethics" taught in Japanese primary and junior high schools. MEXT started to provide free learning materials called *Kokoro no Nōto* in 2002. In 2014, MEXT decided on a major reform of the content of this textbook and printed it again under the title of *Watashitachi no dōtoku*. It is distributed for free to all Japanese primary and junior high schools. The book's edition used in this study is divided into four main chapters, covering topics such as "Living as a member of society" or "Supporting one another".

I used the two-abovementioned resources, one used in schools in Japan the 1930s and one still in use today as input to be analysed with an automatic moral and emotional reasoning agent, in order to verify to what degree what is taught as ethics in school overlaps with the general population's understanding of ethics.

4.3.2 Language resources for application in information processing tools

4.3.2.1 *Aozora Bunko*

Aozora Bunko (青空文庫, literally the "Blue Sky Library", also known as the "Open Air Library")[4] is a Japanese digital library that encompasses several thousands of works of Japanese-language fiction and non-fiction, including out-of-copyright books or works that the authors wish to make freely available.

Aozora Bunko was created on the Internet in 1997 to provide broadly available, free access to Japanese literary works whose copyrights had expired. Most of

[4] https://www.aozora.gr.jp/

the texts provided are Japanese literature, and some translations from English literature. The resources are searchable by category, author, or title. The files can be downloaded in PDF format or simply viewed in HTML format.

In 2013, the Future of Books Fund (本の未来基金 *Hon no mirai kikin*) was established independently to assist funding and operations for *Aozora Bunko* . *Aozora Bunko* currently (as of 8 February 2021) includes more than 16,300 works, a majority of which are novels.

4.3.2.2 Yet Another Corpus of Internet Sentences (YACIS)

Yet Another Corpus of Internet Sentences (YACIS) [55] is the largest Web based blog corpus available for Japanese language, collected automatically from the pages of Ameba blog service and containing 5.6 billion words within 350 million sentences (fFor a precise description of this corpus, please see Chapter 3.3.2).

I used YACIS first as the corpus of contemporary Japanese based upon which I checked for emotional and moral associations with ethical teachings from both the 1930s and today's Japan.

I then performed a second experiment, in which I compared the contents of the older textbook, *Shūshin kyōjuroku*, with a sample created of books published before 1946 included in the Aozora Bunko library to see to which degree what was taught as part of ethics classes in Japanese schools pre-Second World War overlapped with the moral compass of larger society.

4.3.3 Information processing tools

4.3.3.1 Moral Reasoning Agent (MRA)

Moral Reasoning Agent (MRA), was first proposed by [59] and further developed by [28]. The moral consequence retrieval agent was based on the idea of Wisdom of Crowd. In particular, it uses a Web-mining technique to gather consequences of actions applying causality relations, to extract from the Web emotional and ethical consequences of actions found in the input . The agent was previously tested on over 100 ethically significant real-world statements, such as "killing a man", "stealing money", "bribing someone", "helping people" or "saving environment". For a detailed description of the functionning of the Moral Reasoning Agent, please see Chapter 3.3.2.

Through running two datasets created on the basis of the two schoolbooks with the Moral Reasoning Agent, I performed a complete analysis of moral and emotional associations with the ethical teachings deemed useful in the 1930s and nowadays, to see how they stand up to a broader societal standard.

The whole process is explained in more detail in the following section.

4.4 Quantitative Analysis

4.4.1 Initial Data

As a first step in preparing my data sample, I performed an OCR of *Shūshin kyōjuroku* and *Watashitachi no dōtoku* and performed sentence segmentation of both texts. Due to the size of the initial sample, based on both texts, which contained 1,665 sentences, I chose to only compare chapters referring to the same

areas of morality – in this case the sacredness of human life, approaches to human life as a valuable gift, a meaningful life and similar[5].

4.4.1.1 Creating Evaluation Dataset from *Shūshin kyōjuroku*

Upon selecting the chapters in question, I created a list of grammatical endings that might imply that the sentence contains an ethical warning or positive evaluation of a given action. The endings contained grammatical forms such as *ga yoi deshō* (が良いでしょう, "It would be a good thing to…") or *toiu wake desu* (という訳です, "It is so, that…"). There were 172 initial results, split into "negative sentences" (e.g. "It is not…", "You shouldn't", etc.), "positive short" ("Let's…") and "positive long" ("It would be a good thing to…") sentences. I then manually proceeded to remove sentences which brought no ethical meaning to them.

This left me with a sample of 77 sentences which I then evaluated with the Moral Reasoning Agent.

4.4.1.2 Creating Evaluation Dataset from *Watashitachi no dōtoku*

Since there is much less text in this textbook (due to a large number of illustrations as well as putting an emphasis on interaction and thus containing exercises to be filled in by the student), instead of choosing grammatical constructions prone

[5]Sub-chapters: 2 (*Ningen to umarete*, 人間と生まれて, "To be born as a human"), 6 (*Jinsei no shūshi*, 人生の終始, "The beginning and end of one's life") , 37 (*Shisei no mondai*,死生の問題, "Questions related to life and death") from part one and 3 (*Jinsei nido nashi*, 人生二度なし, "You only live once"), 4 (*Seimei no aiseki*, 生命の愛惜, "The caress of life"), 14 (*Jinsei no fukasa*, 人生の深さ, "The depth of life") and 29 (*Jinsei ha myōmi shinshin*, 人生は妙味深々, "Life is mysterious") from part two of *Shūshin kyōjuroku* and chapters 3-1 (*Kakegae no nai jita no seimei wo sonchō shite*, かけがえのない自他の生命を尊重する, "Respecting the irreplacable life of yourself and others"), 3-2 (*Utsukushii mono he no kandō to ikei no nen wo*, 美しいものへの感動と畏敬の念を, "Let yourself be impressed and admire beautiful things") and 3-3 (*Ningen no tsuyosa ya kedakasa wo shinji ikiru*, 人間の強さや気高さを信じ生きる, "Live by believing in human strenght and nobility") from chapter 3 of *Watashitachi no dōtoku*, entitled *Seimei wo kangaeru*, 生命を考える, "Thinking about life"

to containing ethics-related statements, I kept all sentences containing an ethical evaluation, which left me with a total of 52 sentences to run through the Moral Reasoning Agent.

The grammatical information contained in those sentences also differed from *Shūshin kyōjuroku* in the sense that much less was in the form of ready-made teachings delivered by a higher authority and much more often in the form of open questions ("What do you think about..." or "Have you ever given reflection to the question of...").

4.4.2 Challenges in Comparison of Two Textbooks

Shūshin kyōjuroku is a written report of the words of a teacher talking and giving life lessons, very rarely a Q&A. Grammatically, it contains a number of different structures, with a lot of emphasis on "what one ought to do", "The right thing to do" – the duties of a human being, as well as one's relations with those above and below in a hierarchical social structure. A part of the book is devoted to studying the lives of significant individuals influencing Japanese ethical thought, such as Shinran[6] or Johann Heinrich Pestalozzi[7]. Since it is a *compte-rendu* (formal report) of a series of lectures intended specifically for pedagogy students, a large part of the material is devoted to questions related to education and upbringing.

Watashitachi no dōtoku, due to it being published in more recent times, sometimes covers a wholly different range of topics, such as "The benefits and drawbacks of living in an information society" or "Contributing to the world as a real inter-

[6]Shinran (親鸞, 1173-1263) was a Japanese Buddhist monk, the founder of the Jōdō Shinshū sect in Japan

[7]Johann Heinrich Pestalozzi (1746–1827) was a Swiss pedagogue and educational reformer who founded several educational institutions both in German- and French-speaking regions of Switzerland in order to overcome illiteracy

national person while being aware of being Japanese," which are absent from the content of *Shūshin kyōjuroku*. The narration is mostly in the style of open questions ("What do you think would be the right thing to do?") as well as exercises intended to engage the reader ("Imagine your behavior in this situation", "Write down your thoughts upon reading this chapter").

The process of creating two datasets from *Shūshin kyōjuroku* and *Watashitachi no dōtoku* to be analyzed using the Moral Reasoning Agent is shown in Figure 4.1.

4.5 Results and Discussion

4.5.1 First Experiment Results

Upon analyzing the two datasets with the Moral Reasoning Agent, I was left with 22 sentences for which the MRA found enough sample sentences in the corpus to enable it to run a complete analysis[8] of moral and emotional associations (among the results, 8 were based on the sentence sample from *Shūshin kyōjuroku* and the remaining 14 on *Watashitachi no dōtoku*).

A detailed analysis for a sample sentence *hito no inochi wo sukuu* (to save peoples lives) is presented in Table 4.1.

More striking results of the first experiment were summed up in Table 4.2, where upon presenting the original sentences, ethically relevant statements extracted from those sentences and sample sentences containing the same ethically relevant statements found in YACIS, for further clarity I emphasized examples where there were surprising twists in terms of the percentage of positive and negative emotional

[8]Containing, apart from an overall evaluation of moral and emotional consequences, a detailed description of moral («worth praise», «correct action») / immoral («illegal act», «worth reprimanding») associations with the sample term

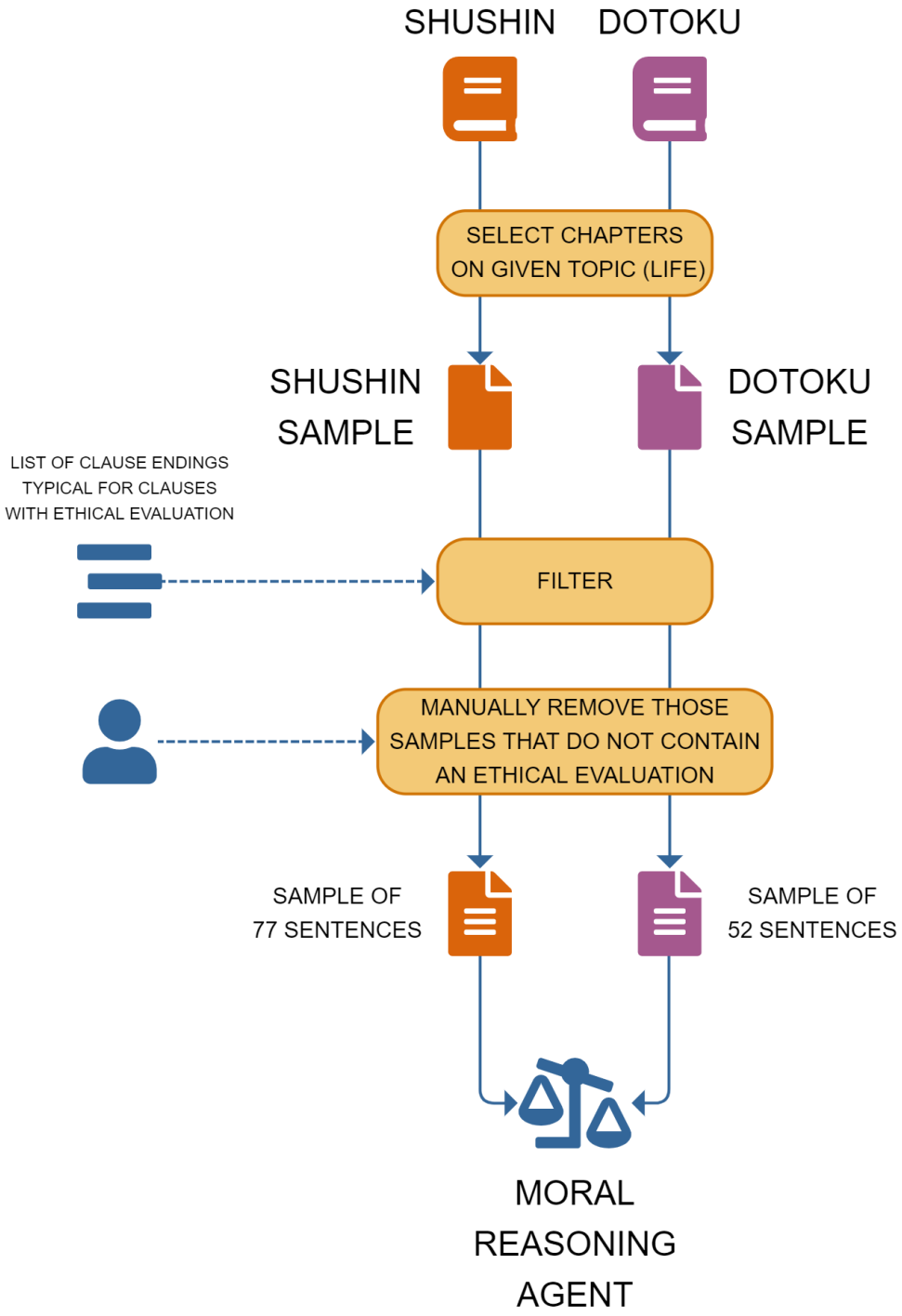

Figure 4.1: Creating two datasets from *Shūshin kyōjuroku* and *Watashitachi no dōtoku* to be analyzed using the Moral Reasoning Agent

	Phrase	*hito no inochi wo sukuu* (“to save people’s lives”)
	Original sentence in *Watashitachi no dōtoku*	人の命を救い、 人々の苦しみを和らげる以外に考えることは何もない。
	Translation	There is nothing to think of except saving people’s lives and alleviating their suffering.
	Sample sentence in YACIS	医療の現場ってとても想像できなくて（特に救命なんてね）、人の命を救うプレッシャーをずっと受け続ける現場なんて、やりたいって思う人なんてめったにいないですよね〜。
	Translation	I can hardly imagine what it feels like to be working in ER (especially saving lives), and rarely do people want to do a job where you are under pressure to save human lives, right?
	Frequency in the corpus	294 times
	Social consequences	positive: 100%, negative: 0%
	Emotional consequences	positive: 50%, negative: 49%
Moral categories	Correct action	23% of the results
	Worth encouraging	14% of the results
	Worth praise	4% of the results
Immoral categories	Illegal act	n/a
	Unacceptable conduct	n/a
	Worth reprimanding	n/a

Table 4.1: Results of the analysis for the term *hito no inochi wo sukuu*

consequences of the actions.

The largest difference in approach in the results of the first experiment is visible when taking into account the emotional value attributed to certain behaviors – with the sample from more recent material giving far more negative or ambivalent responses – which would indicate a high level of awareness of the fact that behaving in the most moral way, beneficial for the broader society is not always pleasant or beneficial to the individual undertaking the action.

4.5.2 Second experiment results

Of course, in order to analyze the mentality of the average Japanese living in the 1930s it would be ideal to compare the sample from *Shūshin kyōjuroku* with a

corpus of sentences from the same time period (namely the 1930s); however, at the time this paper was written the authors were not aware of the existence of such a corpus. I decided to settle for the next best thing, namely to replace the corpus used by the Moral Reasoning Agent and instead of YACIS use texts from Aozora Bunko, a Japanese digital library, published before 1946 and thus influenced by ethical ideas from the 1930s and before. As a first task, I segregated the corpus – since it is not classified according to time periods and there is no search engine that would allow for a robust selection of texts, I chose to retain only works for which the date of first publication was given. I also removed all translations of foreign works, i.e., all items with a foreign title given in the 原題 (*gendai*, "original title") category, of which there were 134 items in total. Finally, I removed items coming from before the Meiji era (1868; 4 items, 2 distinct works). This preliminary data cleaning left us with a total of 1873 texts published before 1946. Through using those texts as my corpus of choice, I enabled the Moral Reasoning Agent to simulate a pre-Second World War moral outlook[9].

This time, out of 58 terms used in the analysis, the Moral Reasoning Agent found enough sample sentences in the corpus to enable it to run a complete analysis of moral and emotional associations for a total of 45 terms. The most remarkable examples of the results of the second experiment were summed up in Table 4.3.

[9]While of course being aware that this solution has certain limitations, as book authors form only one group of citizens, generally better educated and having more liberal opinions than an average member of society. There is also the question of whether the contents of a work of literature represent the opinions of the author

4.5.3 Discussion

In the results of both the first and the second experiment, one can perceive a discrepancy between behaviors beneficial to the individual engaged in a certain action and to broadly perceived society – in terms such as 苦労 (*kurō*, suffering) or正直 (*shōjiki*, honest).

This sharp distinction between understanding what is beneficial to society as a whole and what is beneficial for the person performing a certain positive action is in line with Charles Leslie's Stevenson's and Alfred Jules Ayer's emotivism (the hurrah/boo theory) [65] – a meta-ethical view that claims that ethical sentences do not express propositions but emotional attitudes and the presence of an ethical symbol in a proposition adds nothing to its factual content [6]. As such, while the respondent to certain ethical claims understands their positive value objectively, the overall positivity of the response is reduced due to the consciousness of its plausible negative emotional impact on the individual in question.

In opposition to the view that conscience is the "inner light", Jeremy Taylor [66], a Christian thinker from the 17th century, famously states, that "conscience is, in most men, an anticipation of the opinions of others". Also, William R. Alger [2] similarly claims, that "Public opinion is a second conscience". This directs us to the assumption that conscience can be perceived as an approximate result of the opinions of other people, and thus openly assessing certain actions as morally correct or being "the right thing to do", all the while maintaining a "private ethics" of sorts, is a natural copying mechanism of humans, notwithstanding external circumstances.

This theory proves interesting in the case of more controversial terms such as 国土 (*kokudo*, the territory of one's country, motherland) or 国民 (*kokumin*, nation)

– the complete results for the two terms are shown in Tables 4.4 and 4.5. While giving an inconclusive assessment in terms of emotional consequences, both terms are judged positively when it comes to the benefit to society. Above all, both terms are present in both the textbook for teaching ethics and the literature of the 1930s and before which served as our corpus in this experiment and are also included in the textbook for teaching ethics used in middle schools nowadays, but at the same time are completely absent from a corpus of modern Japanese (and thus one might say from the conscience of the broader population nowadays)[10].

Those results seem to indicate that while subject to a school curriculum centered around values such as nation, serving others, devotion to the motherland, the Japanese chose to superficially believe or attribute value to what they were taught was morally proper in ethics lessons. At the same time, privately, they maintained a more inconclusive approach to several ethical problems. Or is it that they completely lose interest in ethical matters related to strongly patriotic ideals after 1946?

One might state that the reform of education undertaken in 1946 had its biggest effects in this aspect – terms broadly associated with nationalism, while still subject of focus of school-taught ethics, seem to leave the broader society indifferent.

While in terms of core moral values there is little difference to be observed between the moral evaluation of an action between sentences from both *Shūshin kyōjuroku* and *Watashitachi no dōtoku* and YACIS – meaning that what was taught

[10] *Watashitachi no dōtoku* contains the following four subchapters to its Chapter 4 (社会に生きる一員として, *Shakai ni ikiru ichiin to shite*, Living as one of the members of a society): ふるさとの発展のために(*Furusato no hatten no tame ni*, For the developement of one's hometown), 国を愛し、伝統の継承と文化の創造を (*Kuni wo aishi, dentō no keishō to bunka no sōzō wo*, Love the country, inherit the tradition and create culture), 日本人としての自覚を持って真の国際人として世界に貢献したい(*Nihonjin to shite no jikaku wo motte ma no kokusaijin to shite sekai ni kōken shitai*, Contributing to the world as a truly international person while maintaining the awareness of being Japanese), 日本人の自覚を持ち世界に貢献する(*Nihonjin no jikaku wo mochi sekai ni k(o)ken suru*, Contributing to the world while maintaining the awareness of being Japanese)

as morally proper in the textbooks juxtaposes with what the general public finds morally proper[11]. Queries for phrases evolving around the sanctity or value of life always returned positive results when it came to assessing the impact on society (for instance in the case of the query *hito no tame ni tsukusu*, "sacrifice/give yourself to people", for which a detailed analysis is shown in Table 4.6, both a sample based on YACIS – i.e., the way Japanese people assess this action now – and in Aozora Bunko – the way they thought of sacrifice for the sake of others in the 1930s – showed positive social consequences).

This is both heartening (no amount of mistaken moral education motivated by short-term political agenda seems to make people lose their integral sense of what is right and wrong) and at the same time disheartening (a huge rewriting of the basic principles governing mandatory education in the area of ethics seems to have, again, a rather limited impact on the core morals of a society), and seems to confirm the definition of ethics which refers to rules provided by an external source, e.g., codes of conduct in workplaces or principles in religions and morals understood as an individual's own principles regarding right and wrong.

This would provide some insight as to how the Japanese people, fed lessons on the vertical construction of society together with militaristic and nationalistic content all through the 1930s and during the Second World War right until the country's surrender in August 1945 would then shake off their entire education and not even two years later, in May 1947, acclaim the promulgation of a pacifist constitution, proclaiming the equality of all men and renouncing the nation's right to belligerency [15].

[11] Which in turn would suggest that in spite of a huge shift in terms of which ethical values were promoted through school education, the understanding of morality, a "moral compass" of the Japanese people remained largely unchanged

Assuming that the real impact of ethics understood as a set of rules taught throughout the school curriculum seems to have only a limited influence on the morality of an individual, and in light of the reforms undertaken recently by the Japanese government, how should ethics lessons be organized? Is it really impossible to condition a child's moral compass through ethics classes? And finally, is it feasible, as the Japanese Ministry of Education seems to imply, to evaluate and grade a student's ethical outlook in the same way you would grade homework in mathematics?

4.6 Conclusions

The objective of this paper was to demonstrate the challenges facing educators trying to implement ethics as part of the school curriculum, on the example of the contents of textbooks for the subject of "Ethics" in school in pre- and post-war Japan. I was particularly interested in finding out whether changes in ethical education influenced core moral values in humans throughout the century, how much pre-WWII moral attitudes have in common with those of contemporary Japanese, and whether a major reform of the guidelines for teaching the school subject of "Ethics" at school after 1946 has changed the way common people approach core moral questions (such as those concerning the sacredness of human life).

For this purpose, I selected textbooks used in teaching ethics at school from between 1935 and 1937, and those used in junior high schools today (2021) and analyzed what emotional and moral associations such contents generated.

As a result, I found out that, despite changes in the stereotypical view on

Japan's moral sentiments, especially due to historical events, past and contemporary Japanese share a similar moral evaluation of certain basic moral concepts. There is however a large discrepancy between how they perceive some actions to be beneficial to the society as a whole while at the same time being inconclusive when it comes to assessing the same action's outcome on the individual performing them and in terms of emotional consequences. Some ethical categories, assessed positively before the war, while being associated with a nationalistic trend in education, have also disappeared from the scope of interest of post-war society.

The findings of this study support suggestions proposed by others that the development of personal AI systems requires supplementation with moral reasoning. Moreover, the paper builds upon this idea and further suggests that AI systems need to be aware of ethics not as a constant, but as a function with a correction on historical and cultural changes in moral reasoning.

Sample phrase (W – *Watashitachi no dōtoku*, S – *Shūshin kyōjuroku*)	% of positive / negative consequences		Sample sentence from YACIS
	Social	Emotional	
感謝の気持ちを持つ (“feel gratitude”) – S	100% / 0%	57% / 42%	感謝の気持ちを持つことは、習慣のようなところがあり、なれないと、違和感があるかもしれません (“Being grateful is like a habit, and if you don’t develop it, you may feel uncomfortable”)
一人で悩まない (“not suffer on your own”) – S	50% / 50%	51% / 48%	一人で悩まないで誰か身近な人にどんどん話そう。ある程度を超えるとさすがにうざいけど吐き出していい。 (“Don’t worry on your own and talk to someone close to you. At some point, though it might be embarassing, you just have to get it out of your system”)
思いやりを持つ (“be empathic”) – S	85% / 14%	13% / 86%	挫折した時も、痛みが分かる優しさや思いやりを持つために生まれてきたんだよって支えてくれた母。 (“My mother supported me when I was frustrated and explained that I was born with kindness and compassion to understand the pain”)
人を励ます (“encourage people”) – W	66% / 33%	55% / 44%	気持ちも落ち込みがちです。こんなオイラは他の人を励ます資格あるんだろうか。 (“I also tend to be depressed. Makes me wonder whether someone like me is qualified to encourage others?”)
人はいつか死ぬ (“we will all die someday”) – S	33% / 66%	38% / 61%	生と死、仕事に対する誇り、自分を理解してくれる人生のパートナー、色々な人の人生があること、人はいつか死ぬこと。 (“Life and death, pride in one’s work, a partner in life who understands me, the life of various people, and the fact we will all eventually die”)
人の命を救う (“to save people’s lives”) – W	100% / 0%	50% / 49%	医療の現場ってとても想像できなくて特に救命なんてね、人の命を救うプレッシャーをずっと受け続ける現場なんて、やりたいって思う人なんてめったにいないですよね〜。 (“I can hardly imagine what it feels like to be working in ER (especially saving lives), and rarely do people want to do a job where you are under pressure to save human lives, right?”)
人間は弱い (“humans are weak”) – W	33% / 66%	33% / 66%	元々人間は弱いでも世のため人のために生き始めたときどんどんどんどん強くなれるどんどんどんどん美しくなれる人間不思議なるもの (“Originally, human beings are weak, but when they start living for others, they can become stronger and stronger, and become more and more beautiful, which is remarkable”)
人を歓迎する (“welcome people”) – S	0% / 100%	57% / 42%	今日はゲストが多く、２人で１人を歓迎するという感じだったので、負担が大きく大変だったのですが、お腹いっぱい食べることができ、すごくしあわせでした。 (“There were so many guests today and it seemed like two people would welcome one, so the burden was big and it was hard, but I was able to eat a lot which made me very happy”)

Table 4.2: More striking results of the first experiment

Sample phrase from *Shūshin kyōjuroku*	% of positive / negative consequences		Sample sentence from Aozora Bunko
	Social	Emotional	
人生 (“life, human existence”)	62%/37%	47%/52%	今日から新しい自分の人生が始まるのだ、そういううちから強い感情が胸いっぱいに溢れて、家のなかにじっとしていられない気持だった。 (From today, my new life will begin!- and as I said this, my chest was full with strong emotions and I couldn’t stay still in my house).
人間 (“humanity, men”)	48%/51%	36%/63%	ただ自分だけの子にするのではなく、御国の役にたつ人間、りっぱに御奉公のできる武士にしたい。 I do not want it to be only my child, but to become a human being who can serve the country, a samurai who can perform public duties splendidly).
仕事 (“occupation, life’s work”)	65%/34%	43%/56%	そのあいだ縫い針洗濯の手仕事をしても、どんな事をしてでもきっと辛抱しとおしてみせます In the meantime, no matter if you work washing or sewing I will be patient with you).
苦労 (“suffering, effort”)	51%/48%	38%/61%	短い人の一生をそんな取り越し苦労をして更に短くするのは、天命に背く罪とも言うべきじゃ。 (It can be said that it is a sin against the natural order of things to make the short life of a person even shorter by making her endure such suffering).
性欲 (“libido, sexual desire”)	100%/0%	50%/50%	小説家とか詩人とかいう人間には、性欲の上には異常があるかも知れない。 (People such as novelists and poets may well have an abnormal libido).
誠 (”truth, reality”)	51%/48%	35%/64%	夫が決して人の情を解しないやうな人物でないばかりか、日本人としての意気と誠を十分にもつた男だと信じれば信じるほど、こんな些細なことに偏窟な自尊心をみせる気が知れない。 (Not only is my husband one to understand other people’s feelings, but what’s more I believe that he is a man who has enough morale and sincerity as a Japanese person, and I find it hard to believe that such a trivial thing would lead him to demonstrate such mistakenly understood self-esteem).
礼 (“politeness, respectfulness”)	68%/31%	41%/58%	とにかくに等しく恩のあるものならば、一方より礼を言いて一方より礼を言わざるの理はなかるべし (Anyway, if if you are equally endebted to two people, it should not be that you thank one without thanking the other).
正直 (“honesty, telling things as they are”)	66%/33%	38%/61%	正直にいって、そのとき志保は初めて妹に妬みを感じた。 (To be honest, at that time, Shiho felt envious of her sister for the first time).
生命 (“life, creation”)	53%/46%	44%/55%	この法によれば、平民の生命はわが生命にあらずして借り物に異ならず。 (According to this law, the life of a commoner is not equal in value to my life and is no different from an object you would lend).
理想 (“ideal, perfection”)	63%/36%	41%/58%	それだのに、女はおれを高邁な理想主義者だと思つてゐるらしく、なかなか誘惑してくれない。 (Women seem to think that I am a true idealist, and do not try to seduce me).

Table 4.3: Remarkable results of the second experiment

	Phrase	*kokudo* (“motherland”)
	Original sentence in *Shūshin kyōjuroku*	われはひとり人間としてこの地上に生まれたばかりでなく、この日本の国土に生まれたということは、さらに大きな喜びでなくてはならぬと思うのです 。
	Translation	Not only born on this earth as human beings, being born in the land of Japan specifically must be a great joy.
	Sample sentence in *Aozora Bunko*	どうすれば日本の国土に相応し、風景と調和し、無事の日にはこころよい住心地と、たのしい安全感とをあたえるような住宅の群れを作りあげて、いよいよわたしたちの愛惜の念を、深くかつ切なるものにし得るかを考えなければならぬ。
	Translation	How can we create housing suitable for the land of Japan, in harmony with the scenery, and providing a comfortable living environment and a pleasant sense of safety on a plain, uneventful day? I have to think deeply about how I can transform my regrets into something deep and meaningful.
	Frequency in the corpus	205 times
	Social consequences	positive: 58%, negative: 33%
	Emotional consequences	positive: 52%, negative: 47%
Moral categories	Correct action	14% of the results
	Worth encouraging	11% of the results
	Worth praise	n/a
Immoral categories	Illegal act	4% of the results
	Unacceptable conduct	n/a
	Worth reprimanding	17% of the results

Table 4.4: Results of the analysis for the term *kokudo*, motherland

	Phrase	*kokumin* ("nation, one's people")
	Original sentence in *Shūshin kyōjuroku*	諸君らが今日たどりつつある道は、諸君らの自覚の浅深いかんにかかわらず、とにかく生涯を、国民教育のために生きる道であります。
	Translation	The road you are taking today is a lifelong road, regardless of your shallow awareness of it. It is a way to live for national education.
	Sample sentence in *Aozora Bunko*	われわれが死を決するところはそれとは違う、大義を顕彰するということはわれわれ自身の問題ではなく、この国民ぜんぶの系体に関するのだ
	Translation	When we decide to die is different, honoring the cause is not our own problem, but a question regarding the whole of this nation.
	Frequency in the corpus	2500 times
	Social consequences	positive: 61%, negative: 38%
	Emotional consequences	positive: 44%, negative: 55%
Moral categories	Correct action	9% of the results
	Worth encouraging	18% of the results
	Worth praise	n/a
Immoral categories	Illegal act	6% of the results
	Unacceptable conduct	2% of the results
	Worth reprimanding	7% of the results

Table 4.5: Results of the analysis for the term *kokumin*, nation

	Phrase	人のために尽くす [*hito no tame ni tsukusu*, "to serve others/ the people"]
	Original sentence in *Watashitachi no dōtoku*	「我」に引掛っているとは、常に自分の利害を中心にして、人のために尽くすということの分からない人間だ
	Translation	Someone who is an egoist is a person who always focuses on his own interests and doesn't understand that one should serve others.
	Sample sentence in YACIS	悲しみと苦痛は、やがて"人のために尽くす心"という美しい花を咲かせる土壌だ
	Translation	Sorrow and pain are the soil that will eventually bloom with beautiful flowers called "serving other people".
	Frequency in the corpus	213 times
	Social consequences	positive: 100%, negative: 0%
	Emotional consequences	positive: 47%, negative: 52%
Moral categories	Correct action	7% of the results
	Worth encouraging	n/a
	Worth praise	n/a
Immoral categories	Illegal act	n/a
	Unacceptable conduct	n/a
	Worth reprimanding	n/a
	Original sentence in *Shūshin kyōjuroku*	すなわち諸君たちが教育を通して国家社会に尽くす
	Translation	In other words you young people will contribute to the nation through education.
	Sample sentence in *Aozora Bunko*	父子二人が身を捧げ、人の為にも尽くす
	Translation	Both father and son will devote themselves and serve others.
	Frequency in the corpus	71 times
	Social consequences	positive: 66%, negative: 33%
	Emotional consequences	positive: 19%, negative: 81%
Moral categories	Correct action	n/a
	Worth encouraging	16% of the results
	Worth praise	n/a
Immoral categories	Illegal act	n/a
	Unacceptable conduct	n/a
	Worth reprimanding	5% of the results

Table 4.6: Detailed analysis for the term *hito no tame ni tsukusu*

Chapter 5

A Case Study on Authorship Analysis of Texts by Arata Osada

5.1 Introduction

Authorship analysis is the process of determining the authorship of a document based on its characteristics. The problem itself has a long history, dating back to the end of the 19th century when Mendenhall [36] for the first time examined the word length in the works of Bacon, Shakespeare and Marlowe in order to detect quantitative stylistic differences.

In a computational context, it is an emerging area of research associated with applications in literary research, cyber-security, forensics, and social media analysis. Other applications include:

- **Forensic linguistics and online bullying detection** (see [12, 17]) – to identify characteristics of the author of anonymous, pseudonymous or forged text, based on the author's use of the language (blackmailing letters, confes-

sions, testaments, suicide letters).

- **Bot detection** – in the context of marketing, social bots can artificially inflate the popularity of a product by posting positive reviews. Especially Twitter bots can be considered a threat given their commercial, political and ideological influence, exemplified for instance by the 2016 United States Presidential Election [11], during which they polarised political conversations, and spread fake news.

- **Marketing** – to identify the demographics of people that like or dislike certain products based on the analysis of blogs, online product reviews and social media content (see [24]).

Much research has been focused on determining suitable features for modeling writing patterns from authors. Reported results indicate that content-based, as well as style-based features continue to be the most relevant and discriminant features for solving this task (see [62]).

The remainder of this chapter is organized as follows. In the following section (5.2), I introduce our research questions and present the background of our study, including previous research on authorship analysis. I also discuss the influence of historical events on changes in ethical values and the cultural importance of the *shinpoteki bunkanin*, a term used to describe Japanese intellectuals active both before and after the Second World War, in Japan. I then introduce the person of Arata Osada, our object of analysis. In Section 5.3, I introduce the resources applied to perform this study. Section 5.4 describes the classification model I applied to our task, as well as the procedure of data generation. Finally, I present the results of my experiments and conclude the paper with a discussion on the

results and some ideas for future work.

5.2 Research questions, previous research, background of the study

In the previous chapter, I investigated whether it is possible to automatically determine if changes in ethical education influence core moral values in humans throughout the century, on the example of Japan (see also [44]). As a result, I found out that, despite the changes in the stereotypical view on Japan's moral sentiments, as well as the changes in the model of education, especially due to historical events (pre- and post-World War II), past and contemporary Japanese shared a similar moral evaluation of certain basic moral concepts. This was an interesting discovery, since it showed that Japan, known for its imperialist and militarist ethical profile before the war, which greatly influenced people's attitudes towards death (e.g., suicide airplane pilots *kamikaze*), in fact had a set of core moral values, which included the importance of preservation of life, which it shared throughout the history despite its official profile as a country.

In this chapter I build upon the above findings. In my work so far I focused on factors influencing the ethical values at a macro-scale of nation-wide population, such as general ethical education policies. In this chapter, on the other hand, I focused on analyzing the factors influencing not a global population, but rather an individual's ethical values.

In particular, I wanted to know if – given that education policies can only influence someone's ethical outlook to a limited extent – it is possible to specify other external stimuli, such as major historical events (e.g., the World War II),

that would change the ethical profile of particular individuals.

The specific research question I set up to answer in this research was twofold. Firstly, since the methods used for performing authorship analysis imply that an author can be recognized by the written content he or she creates (Internet messages, articles, books, etc.), how accurate would they prove in the case of an author who diametrically changed their opinions due to an external stimulus, such as the impact of war?

From the point of view of authorship analysis, I wanted to know if it would be possible for an authorship analysis solution to correctly attribute works from before and after the war to the same author and with what accuracy. Another thing I checked was whether such a shift in opinions could be precisely quantified, and on what level it could be perceived. Moreover, I checked how would such a difference compare to the situation where two completely different authors are taken into account.

From the point of view of the impact on a person's ethical outlook as expressed in their writings, if the content of an author's work changed in between before and after the war, and if the authorship analysis system encounters difficulties in detecting single authorship, it would mean that historical events can have a great impact on a person's ethical outlook.

5.2.1 Previous research

Authorship analysis is the task of examining the characteristics of a document in order to draw conclusions about its authorship [72].

A 1964 study by Mosteller and Wallace [42] on the authorship of "The Federalist Papers" (a series of 146 political essays written by John Jay, Alexander Hamilton,

and James Madison, whose authorship was controversial), where Bayesian statistical analysis of word frequencies was used, initiated non-traditional authorship analysis studies, no longer relying on a human specialist to determine authorship.

Until the ascension of the Internet and social media, research in authorship analysis was dominated by a computer-assisted, but not computer-based approach relying on the process of identification on quantifiable language features, such as word length, phrase length, sentence length, vocabulary frequency, distribution of words of different lengths, etc. [21], known as *stylometry*. The limitations of this approach included, among others, a small number of candidate authors being taken into consideration and the lack of suitable benchmark data [64].

The propagation of the Internet and social media and the virtually immeasurable amount of electronic texts available through it proved to be a great stimulus for the evolution of Natural Language Processing, Information Retrieval and Machine Learning, and as a consequence authorship analysis solutions as well. At the same time, the amount of text information to process and categorize indicated the potential of authorship analysis in several applications [34] and the need for a reliable, computational method to perform it. In the last decade, areas of research in connection to authorship analysis include efforts to develop practical applications dealing with real world texts rather than solving literary questions.

In the typical authorship analysis problem – known as closed-set authorship identification or authorship attribution – a text of unknown authorship is assigned to one candidate author, given a set of candidate authors for whom text samples of undisputed authorship are available. In open-set attribution, on the other hand, the true author is not necessarily included in the set of candidate authors. A special case of open-set attribution is authorship verification, where, given one or more

documents by a single author and another, anonymous document, the task is to determine if that document was also written by the same author [52, 29].

Existing approaches to authorship analysis can be divided in two main groups: similarity-based methods and machine learning-based methods [29].

Recently, a number of studies have been carried out on cross-domain authorship identification [26] (where the texts of known and unknown authorship belong to different domains) and style change detection (where single-author and multi-author texts are to be distinguished), featuring several methods involving the use of n-grams [61] and deep learning [8, 56, 39]. Nirkhi et al. [45] investigated the effect of increasing the number of authors on an SVM-based authorship identification system's performance.

Azarbonyad et al. [7] analyzed the changes in word usage by authors of tweets and emails and proposed a similarity-based, time-aware authorship attribution approach.

I am not aware of any previous research investigating the effect of temporal changes in the context of machine learning-based authorship analysis.

Outside of the domain of automatic authorship attribution systems, Rexha et al. [58] conducted a study to determine if human evaluators can identify authorship among texts with high content similarity, and what features influence their decisions.

Concerning research involving the use of pre-trained language models (such as BERT) in authorship analysis, Barlas et al. [9] extended the successful authorship verification approach of Bagnall [8], based on a multi-headed classifier, by combining it with four different types of pre-trained language models. Most recently, Shimizu [63] reported the results of an authorship identification experiment for Japanese, using data obtained from the Aozora Bunko (an open-source repository of Japanese

literature, also used in this research) and BERT.

5.2.2 Influence of historical events on change in ethical values

As has been discussed in political studies, such as the one by Maja Zehfuss, [70], ethics, especially in the form of specifically manipulated codes of ethics taught to people, often become a strong motivation for war. In such cases, ethical considerations do not act as a constraint – on the contrary, making a commitment to ethics enables war and enhances its violence. However, war is also an agent of major changes – wars may change individuals' value systems and influence future choices, as, for example, when individual war experiences shift their evaluation of costs and benefit. War-weariness (and by consequence a sharp turn towards pacifism) is one such effect. Finally, wars may entail structural changes for actors and unchosen shifts in the context or environment within which they act[1]. I chose the Second World War as a cut-off point in this research due to its major impact on the Japanese society as a whole and the Japanese intelligentsia[2] specifically. Especially the Japanese intellectual elites, as mentioned by John Dower [15], "performed a virtuoso turnabout, since only a precious few opposed the war and virtually any author wanting to be published in the 1930s had to be a military enthusiast and a supporter of the idea of the expansion of Japan in Asia."

[1] http://www.grandstrategy.net/Articles-pdf/evaluating_war.pdf
[2] Artists, teachers, academics, writers, etc.

5.2.3 Cultural importance of *shinpoteki bunkanin* in Japan

Shinpoteki bunkanin (進歩的文化人, "progressive men of letters") is a term used to describe Japanese intellectuals active after the Second World War, who were instrumental in disseminating democratic ideas and pacifism in Japan. However, the same people often supported Japanese militarism in the 1930s and quickly changed sides upon Japan's defeat, often trying to conceal or destroy their pre-war publications, of which many reflect their peculiar shift of political opinions.

For this reason, to analyze how a major life event like war influences one's ethical values I chose one of the representatives of *shinpoteki bunkanin*, namely Arata Osada, whose literary works made a clear and major turn after the war, from a militarist to an extreme pacifist attitude. Since many literary works of pre- and post-war writers are freely available as a language resource, it is possible to quantify how exactly such attitudes changed after the war. To perform such quantification, in this research I apply a neural language model-based method for single authorship identification.

5.2.4 The person of Arata Osada: object of analysis

Arata Osada (長田新, 1887-1961) was a Japanese educator, honorary professor of Hiroshima University and specialist in the history of education. Before and during the Second World War he became known for his vocal views on patriotic education based on the German model (*Nationalpädagogik* known in Japanse as *Kokka kyōikugaku*, 国家教育学, or "National education", 1944) as well as newspapers articles with a militaristic undertone[3]. On August 6, 1945, the atomic bomb was

[3]An example excerpt: "War is the motivation for the advance of humanity. Japanese want to be reborn 7 times to serve their country and this is so out of their great love for the motherland. The Japanese army is powerful because each of them wants to be like a living shield to the

dropped on Hiroshima, in which attack Osada was seriously injured. In 1947, he became the first chairman of the Japanese Educational Research Association and was a professor at the University of Hiroshima until his retirement in 1953. He became one of the key players in post-war Japanese education reconstruction, including forming the Japanese Children's Association and serving as its first president. Based on his experience of the atomic bomb, Osada actively participated in the peace movement opposing nuclear arms, and collected the notes of boys and girls who experienced the dropping of the atomic bomb and published it as *Genbaku no ko – Hiroshima no shōnen shōjo no uttae* (原爆の子〜広島の少年少女のうったえ, "Children of the Atomic Bomb – The Plight of Boys and Girls of Hiroshima"). He made a public appeal to abolish the imperial system in the magazine *Kyōiku* (教育, *Education*, in September, 1956), saying "[...] abolish the imperial system, the defeat in the war was a win in this sense – it will become a path to democracy". His other post-war publications include *Heiwa wo motomete* (平和を求めて, "In search for Peace", 1962) and *Shakaishugi no bunka to kyōiku – watakushi no mita Soren to Chūgoku* (社会主義の文化と教育 わたくしのみたソ連と中国, "Socialist culture and education – the Soviet Union and China as I saw it", 1956).

I chose his works as a basis to create our sample in this research due to the fact that most of his writing is centered around the subject of education of the youth (thus the topical content or spectrum remain unchanged in pre- and post-war publications), but the consequences of the Second World War seem to have brought on a major shift in his belief system – from loyal imperialist and militarist to an extreme pacifist.

Emperor and sacrifice their lives. Perfecting the army's education is like perfecting one's life [...]. Soldiers are determined to sacrifice their lives without regret for one "absolute person" [i.e., the emperor]" – excerpts from an article in the journal *Seishonen Shidō* (青少年指導, "Instructing of youth and children") from February 1944.

5.3 Materials

Aozora Bunko Aozora Bunko[4] (青空文庫, literally the "Blue Sky Library", also known as the "Open Air Library") is a Japanese digital library that encompasses thousands of works of Japanese-language fiction and non-fiction, including out-of-copyright books or works that the authors wish to make freely available see detailed description in previous chapter 4, subsection 4.3.2.1.

Books by Arata Osada As the main test data I used four books authored or co-authored by Arata Osada, two among which were written before the Second World War: *Kyōiku shisōshi* (教育思想史 [47], "The history of educational thought"), published in 1931, and *Shinkyōiku no kōsō – Amerika no bunka-kyōiku wo hihan shite* (新教育の構想 アメリカの文化・教 育を批判して [49], "The making of a new education – Critical thoughts on American culture and education"), written before and during the war and published in 1949; and two others written after the war: *Nihon no unmei to kyōiku* (日本の運命と教育 [48], "Japan's destiny and education"), published in 1953, and *Kyōiku kihonhō* (教育基本法 [46], "Basic Law of Education"), published in 1957. I chose those four books due to the common topic that they cover, namely education.

As a first step in preparing my data sample, I performed an OCR of the four above-mentioned books and in the case of books co-authored by Osada, selected only the fragments that he had authored.

[4] https://www.aozora.gr.jp/

5.4 Same authorship detection system

I set out to answer our research questions through performing an authorship analysis experiment, using a binary text classification model based on a pre-trained language model (BERT), trained to predict if two fragments of text presented to it were produced by the same author or two different ones.

Recent years have witnessed major improvements in a number of Natural Language Processing benchmarks, owing to the advent of deep pre-trained language models (examples include text classification [69], text summarization [71, 68], question answering [50, 14] and machine translation [31]). One of them is BERT (Bidirectional Encoder Representations from Transformers), proposed by Devlin et al. [14]. One of the two tasks used in pre-training of a BERT model is next sentence prediction, where the model learns to predict whether two sentences are likely to occur next to each other in a corpus, or are unrelated. This makes BERT a natural choice for my problem, which is similar.

I employed the Japanese version of BERT, released by the Tohoku University's Inui Laboratory[5] (specifically, the variant using both MeCab and WordPiece tokenization, without whole word masking). The model was pre-trained on Japanese Wikipedia articles.

5.4.1 Data

The training set for my classifier consists of randomly picked samples from the Aozora Library, where each sample is made of two paragraphs, separated by a special token (*[SEP]*). 50% of the data set is comprised by positive samples (i.e.,

[5] https://github.com/cl-tohoku/bert-japanese

those where both paragraphs have the same author), with the label set to "1". The remaining half are negative samples, with the "0" label. Positive samples are further divided in three sub-categories of equal size:

- two paragraphs from the same document;
- two paragraphs from different documents by the same author, where the time lapse between the publication of the first and the second document is less than or equal to 10 years;
- two paragraphs from different documents by the same author, where the time lapse between the publication of the first and the second document is more than 10 years.

Negative samples are subdivided in two equally sized parts:

- paragraphs from two documents by different authors, where the time lapse between the publication of the first and the second document is less than or equal to 10 years;
- paragraphs from two documents by different authors, where the time lapse between the publication of the first and the second document is more than 10 years.

I chose 10 years of time lapse as the cutting point in this experiment due to it being approximately the time between writing the last book in the pre-war sample and the first post-war book by Arata Osada[6].

[6]While it was published in 1949, *Shinkyōiku no kōsō – Amerika no bunka-kyōiku wo hihan shite* represents Osada's pre-war, nationalist beliefs. For this reason, in my experiment we treated it as a book from 1940, which is the estimated time of its writing. For all other books in my data, I took into account the year of the first publication.

Due to the maximum sequence length of 512 sub-word tokens, imposed by the pre-trained BERT model used in the experiment, longer paragraphs were truncated to keep the combined token count of both paragraphs in each sample within the limit. In order to avoid significant differences in length between two paragraphs constituting a data point – which might affect the system's performance – I filtered out paragraphs with the number of tokens smaller than 200. This left me with a total of 66,447 paragraphs in 6,111 documents by 412 authors[7]. Authors were then split in three groups: (i) authors with only a single document, (ii) authors with multiple documents published within a period of 10 years and (iii) authors of multiple documents published over a period longer than 10 years. After that, each group was randomly divided between the training set, development set, and test set, proportionally to the size of each data set. Finally, I generated a specified number of data points for each data set and category of samples, by randomly sampling pairs of authors, their documents and paragraphs they comprise.

The test set composed of the four books by Arata Osada was compiled according to the same rules, the only difference being the fact that at least one of the paragraphs in each data point was sampled from his works (in the case of negative samples, the other paragraph was picked from the same pool of documents as those used in the Aozora Bunko test set). Furthermore, I applied the same guidelines to create separate test sets for three other individual authors (excluded from the Aozora Bunko data set), with the aim of using them for direct comparison with Osada. These were: Yukichi Fukuzawa (福沢諭吉, 1835-1901), an educator, entrepreneur, political scientist and translator of foreign literature[8], Kyōka Izumi

[7]These statistics do not include three authors (namely, Yukichi Fukuzawa, Kyōka Izumi and Asajirō Oka), whose works were excluded from the Aozora data set and used to build separate, single-author test sets for direct comparison with Arata Osada.

[8]I included his works: *Gakumon no susume*(学問のすすめ, 1872), *Gakumon no dokuritsu* (学

(泉鏡花, 1873-1939), a novelist writing mostly about societal matters[9] and Asajirō Oka (丘浅次郎, 1868-1944), a researcher writing about natural history[10].

Since my main interest was in investigating an authorship analysis system's performance when applied to two different documents by the same author (such as Arata Osada), samples where both snippets originate from the same document were not indispensable for the training of the classifier. At the same time, I expected that reserving more capacity in the training set for the other two types of positive samples may result in improved accuracy. To verify that hypothesis, in an additional experiment I trained a model on data without same-document samples.

In order to minimize the potential effect that the characteristics of a particular random sample extracted from my data might have on the experiment results, I generated all data sets three times, each time with a different author split. In Section 5.5, I report the results calculated from the sum of predictions made by models trained and tested on all three variants.

Tables 5.1 and 5.2 show statistics of data sets used in both experiments. A sample data point from the training set is shown in Table 5.3. During training, the model was only presented with the last two fields: "Paragraph1+2" and "Label".

問の独立, 1873), *Shōgaku kyōiku no koto*, (小学教育の事, 1879), *Tokuiku ikan* (徳育如何, 1882), *Gakumon no dokuritsu* (学問の独立, 1883), *Dokurinri kyōkasho* (読倫理教科書, 1890), *Onna daigaku hyōron* (女大学評論, 1899), *Shin onna daigaku* (新女大学, 1899) and *Shūshin yōryō* (修身要領, 1900).

[9]I included a novelist in order to test my assumption that in the case of a fiction writer, covering a broad range of topics in between their works, the performance of the model would be worse than in the case of an author writing on one topic throughout their works. I included: *Katsu ningyō* (活人形, 1893), *Giketsu kyōketsu* (義血侠血, 1894), *Yakōjunsa* (夜行巡査, 1895), *Bakeichō* (化銀杏, 1896), *Ryūtandan* (竜潭譚, 1896), *Kaijō hatsuden*, 海城発電, 1896), *Yushima mōde* (湯島詣, 1899), *Sanmai tsuzuki* (三枚続, 1900), *Kōgyoku* (紅玉, 1913), *Nihonbashi* (日本橋, 1914), *Hinagatari* (雛がたり, 1917), *Shippō no hashira* (七宝の柱, 1917), *Osaka made*(大阪まで, 1918) and *Hakushaku no kanzashi*(伯爵の釵, 1920).

[10]I used: *Dōbutsu sekai ni okeru zen to aku* (動物界における善と悪, 1902), *Jinrui no kodaikyō* (人類の誇大狂, 1904), *Shizenkai no kyogi* (自然界の虚偽, 1907), *Seibutsugakuteki no mikata* (生物学的の見方, 1910), *Warera no tetsugaku* (我らの哲学, 1921) and *Ningenseikatsu no mujun* (人間生活の矛盾, 1926).

	Aozora train.	Aozora dev.	Aozora test	Y. Fukuzawa	K. Izumi	A. Oka	A. Osada
Authors	356	22	34	35	35	35	35
Documents (mean)	4395	343	312	224	242	235	252
Data points	Same author & document						
	6,000	400	600	600	600	600	600
	Same author, diff. documents, dist. <= 10 y.						
	6,000	400	600	600	600	600	600
	Same author, diff. documents, dist. > 10 y.						
	6,000	400	600	600	600	600	600
	Different authors, distance <= 10 years						
	9,000	600	900	900	900	900	900
	Different authors, distance > 10 years						
	9,000	600	900	900	900	900	900

Table 5.1: Statistics of data sets used in the experiment with all 5 types of samples. Since the numbers of documents in each of the 3 variants differ, I report the average values.

5.4.2 Model training

To train the classification system, I fine-tuned the Japanese BERT model on the Aozora training set for one epoch (in preliminary experiments I fine-tuned it for up to 5 epochs, but no further improvements were observed on the development set). I used a learning rate of 3e-6 and batch size of 16. The classifier was implemented with the Flair library[11] [1].

As explained in the previous section, all data sets were created in three different permutations. Furthermore, since I noticed a fair amount of variability in the validation results between multiple training runs on the same data, I repeated the training process 5 times on each variant of the data. As a result, I trained and

[11] https://github.com/flairNLP/flair

	Aozora train.	Aozora dev.	Aozora test	Y. Fukuzawa	K. Izumi	A. Oka	A. Osada
Authors	356	22	34	35	35	35	35
Documents (mean)	4597	365	323	219	248	242	255
Data points	Same author, diff. documents, dist. <= 10 y.						
	9,000	600	900	900	900	900	900
	Same author, diff. documents, dist. > 10 y.						
	9,000	600	900	900	900	900	900
	Different authors, distance <= 10 years						
	9,000	600	900	900	900	900	900
	Different authors, distance > 10 years						
	9,000	600	900	900	900	900	900

Table 5.2: Statistics of data sets used in the experiment without same-document samples. Since the numbers of documents in each of the 3 variants differ, I report the average values.

tested a total of 15 models – in Section 5.5, I report the results calculated from the sum of predictions made by all of them.

5.5 Experiment results

The results of the main experiment are summarized in Table 5.4 and Figures 5.1 and 5.3. In Figure 5.1, the bottom row of each graph represents the results on the test set made based on Aozora Bunko, while the four other rows represent the results on the samples based on works of Arata Osada and the three authors from the comparison sample. The blue bar represents the accuracy with which the model was able to identify authorship, while the orange bar represents the confidence with which the choice was performed (i.e., softmax probability of the predicted class).

In the first graph (Fig. 5.1.1), I present the results of single authorship identifi-

Author1	000525
Year1	1933
Document1	43239
Paragraph1	「先づ今日の時勢よりお話申しますと、世人の社会を見ることが、簡単過ぎて居ると思ひます。 [...] また政治の領分は、社会的事物の大なる者だが、社会全体から見ればその一部分である。されば仮令政治の弊害全部を破りました所で、社会の一部が良くならうが、全部は良くなりませぬ。
Author2	000525
Year2	1904
Document2	3507
Paragraph2	「左様です、彼は決して嫉妬などの為めに凶行に出でたのではありません、 [...] 又た其の高潔なる愛情の手に倒れたと云ふことは、女性としての満足なる生涯では無いでせうか」
Paragraph1+2	「先づ今日の時勢よりお話申しますと、世人の社会を見ることが、簡単過ぎて居ると思ひます。欧羅巴の人々も過去の歴史に於て、矢張りこの過ちを重ねて居ります。それは何かと言ふに、政治上目に見える弊害があれば、その弊だけを止めれば、社会は大層良くなると思うて、尽力して弊を除きましても、社会は思つた程良くならない。こゝに於て大失望して大騒動となる。我国に於きましても、政治上に種々なる弊害があるから、欧羅巴の様に改革したならば、定めし黄金世界になるであらうと思つて居る者があるが、これは余りに政治に重きを置き過ぎたものと、私は考へます。社会を組立てるには色々道具がある。家族であるとか、国民の教育であるとか、これを集めて社会が出来て居る。この社会的原素の中に政治と云ふものがある。また政治の領分は **[SEP]** 「左様です、彼は決して嫉妬などの為めに凶行に出でたのではありません、——必竟、自分の最愛の妻——仮令結婚はしないにせよ——を、姦淫の罪悪から救はねばならぬと云ふのが、彼の最終の決心であつたのです、彼の此の愛情は独り婦人に対してのみで無いのです、彼が平生、職業に対し、友人に対し、事業に対する観念が皆な其れでした、成程、其の小米と云ふ婦人も、今ま貴女の（と花吉を一瞥しつ）仰つしやる通り実に気の毒でした、然かし彼女が彼の如くして生きて居たからとて、一日と雖も、一時間と雖も、幸福と云ふ感覚を有つことは無かつたでせう、兼吉が執つた婦人に対する最後の手段は、無論正道をば外れてたでせう、が、生まれて此の如き清浄な男児の心を得、又た其の高潔なる愛情の手に倒れた
Label	1

Table 5.3: Sample record from the training data

cation on a combination of two fragments from the same document. In this case, the system performed well above the average, with accuracy exceeding 90% on all but one test set and 99% on two of them. This means that apart from being able to capture the relationship between adjacent sentences (as in the next sentence prediction objective), BERT is also capable of discerning similarities in content between distant samples from the same work.

The second graph (Fig. 5.1.2) demonstrates how the model performed when presented with paragraphs of texts written by the same author, but coming from different books, written in the space of less than or 10 years. The third graph (Fig. 5.1.3) is where results are given for a setting with paragraphs of texts written by the same author, but coming from different books, written in the space of more than 10 years. In the first case, a drop in accuracy of 18.1% on average was observed, and increasing the distance in years resulted in a further drop of 9.6%.

In all three categories of positive samples, results for Kyōka Izumi were the lowest, which is consistent with my hypothesis that a diversified sample such as excerpts of novels with different plots and subjects and thus diverse content would affect the performance of the model. It is also confirmed by the fact that the results on the Aozora Bunko set – consisting largely of fiction works – were the second worst. Surprisingly this tendency is also visible in the case of samples created from a single document, implying that novels are more diverse in terms of content also within a single document and as such represent more of a challenge for authorship analysis systems.

Figure 5.2 and Table 5.5 present the results yielded by the classifier trained on data without same-document samples and a comparison with the model trained on all five types of samples. While it did correctly detect same authorship for a

greater number of samples from the remaining two categories of positive samples (presumably due to increased capacity in the training data), it was at the cost of noticeably lower F-score for different-author samples. This suggests that training samples where both fragments of text originate from the same document are beneficial in the modeling of other categories, as well.

Test set	Class	Precision	Recall	F_1 score
Arata Osada	1	0.818	0.807	0.813
	0	0.810	0.821	0.815
Asajirō Oka	1	0.817	0.865	0.840
	0	0.856	0.806	0.830
Kyōka Izumi	1	0.783	0.590	0.673
	0	0.671	0.836	0.745
Yukichi Fukuzawa	1	0.806	0.971	0.881
	0	0.963	0.766	0.853
Aozora Bunko	1	0.791	0.736	0.763
	0	0.753	0.805	0.778
OVERALL	1	0.804	0.794	0.799
	0	0.796	0.807	0.802

Table 5.4: Experiment results.

Based on the experiment, what answer should be given to my main research question: did the change in Arata Osada's ethical values reflect itself in the predictions yielded by the authorship analysis system? While – as shown in Fig. 5.1.2 and 5.1.3 and Fig. 5.3 (first column) – there was a significant drop in accuracy as the distance in years between two documents increased, the same was also true for other authors (with the exception of Asajirō Oka) and the Aozora Bunko data set. The fact that the results on test data based on fiction works (i.e., Kyōka Izumi's novels and the Aozora Bunko data set) were consistently worse than those for Osada's books, seems to indicate that a change in the author's opinions is less relevant to the model than a change in topic (such as between two novels

by the same author, with different characters and/or settings). Furthermore, I did not observe a clear-cut difference between the results on test samples composed exclusively of fragments from pre-war or post-war books and those combining texts from both periods: the accuracy for samples including one paragraph from each of Osada's two pre-war books (i.e., those where the distance in years is 9 – see Fig. 5.3) was lower than for samples combining the *Shinkyōiku no kōsō – Amerika no bunka-kyōiku wo hihan shite* with fragments from either of the post-war books (i.e., those where the time lapse is 13 or 17 years).

On the other hand, in terms of accuracy for samples comprising paragraphs from distant documents, the results on Osada test set were substantially lower than in the case of other non-fiction writers (i.e., Yukichi Fukuzawa and Asajirō Oka). Interesting conclusions can also be drawn from observation of the changes in probabilities assigned by the classifier to its predictions. While in the case of Asajirō Oka and Yukichi Fukuzawa, a strong correlation exists between the confidence scores and accuracy, on the remaining three data sets – including the Osada test set – there is a widening gap between the two values as the time lapse increases (see Fig. 5.3), which means that the model was confident that texts written by the same author were actually written by two different people. This leads me to believe that a change in opinions as expressed in writing can affect authorship analysis. If before and after experiencing a World War the opinions that Arata Osada formulated in his writings changed enough for the model to encounter difficulties while trying to detect single authorship, this might also mean that the impact of historical events, such as a war, on a person's ethical outlook is significant enough to make them express themselves in writing like a different person or in other words "make them a different person".

The fourth (Fig. 5.1.4) and fifth (Fig. 5.1.5) graph and the second column of Fig. 5.3 illustrate the results when the classifier was presented with two fragments of text written by different authors. Contrary to same-author samples, in this case the performance tends to increase with the distance between the dates of publication, since the style of writing and language evolves over time, making it easier for the model to make a distinction between authors and come to the conclusion that samples provided are authored by two different people.

		Model trained with same-document samples:					
		NO			**YES**		
Test set	**Class**	**Prec.**	**Rec.**	**F_1**	**Prec.**	**Rec.**	**F_1**
Arata	1	0.775	**0.772**	**0.773**	**0.805**	0.719	0.760
Osada	0	**0.773**	0.776	0.774	0.746	**0.826**	**0.784**
Asajirō	1	0.763	**0.824**	0.792	**0.807**	0.796	**0.802**
Oka	0	**0.808**	0.744	0.775	0.799	**0.810**	**0.804**
Kyōka	1	0.707	**0.547**	**0.617**	**0.738**	0.470	0.574
Izumi	0	**0.631**	0.773	0.695	0.611	**0.834**	**0.705**
Yukichi	1	0.773	**0.968**	0.860	**0.808**	0.956	**0.876**
Fukuzawa	0	**0.957**	0.716	0.819	0.947	**0.773**	**0.851**
Aozora	1	0.723	**0.694**	**0.708**	**0.765**	0.635	0.694
Bunko	0	**0.706**	0.734	0.720	0.688	**0.805**	**0.742**
OVERALL	1	0.752	**0.761**	**0.756**	**0.790**	0.715	0.751
	0	**0.758**	0.749	0.753	0.740	**0.810**	**0.773**

Table 5.5: Comparison of the results achieved by models trained with and without same-document samples, on test data without such samples (best results in bold).

5.6 Discussion

Currently developed solutions in the area of automatic authorship analysis are based on the assumption that there are certain stable, unchanging factors or features to the content an individual generates, on which the whole process of identification

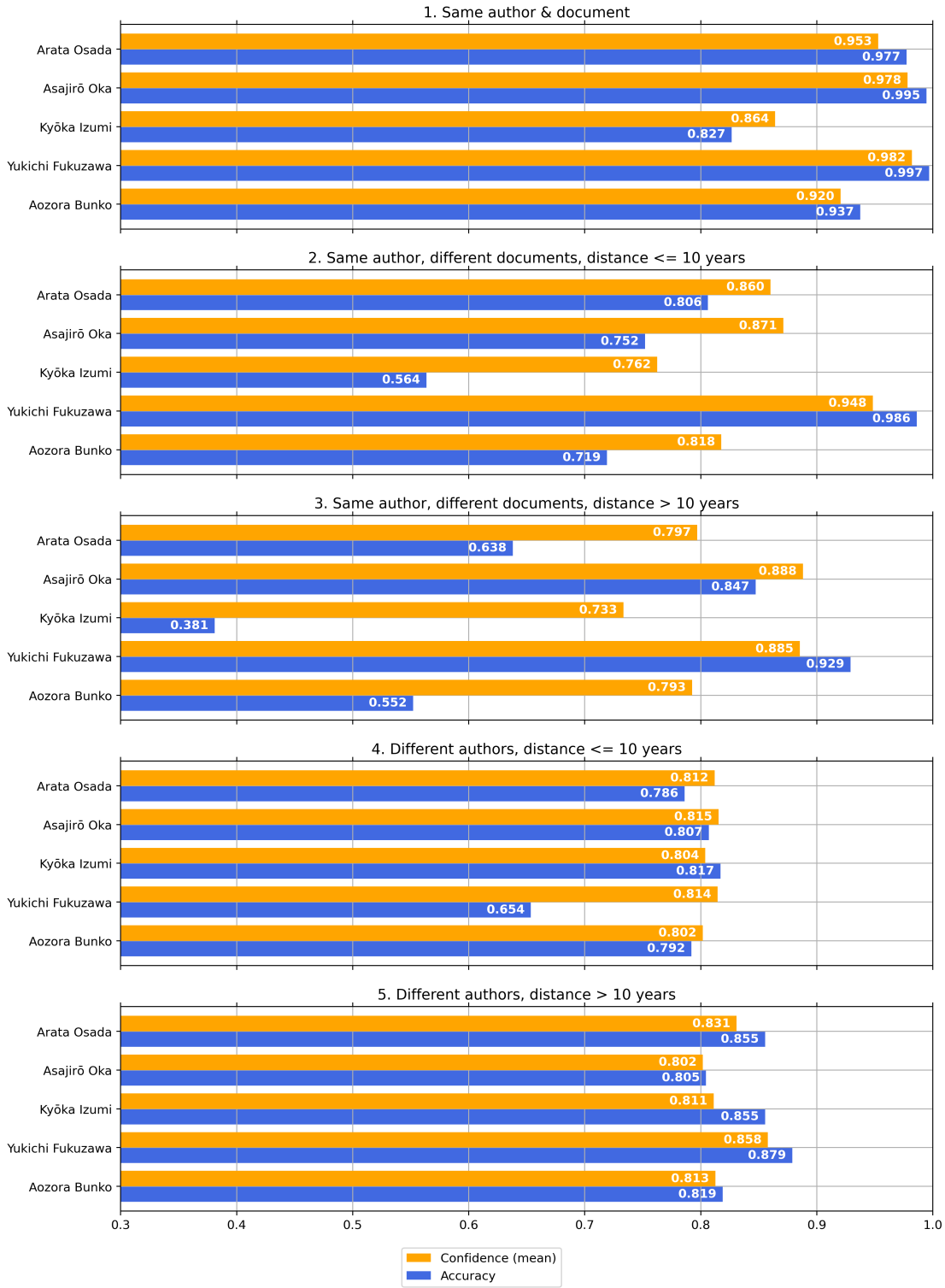

Figure 5.1: Experiment results.

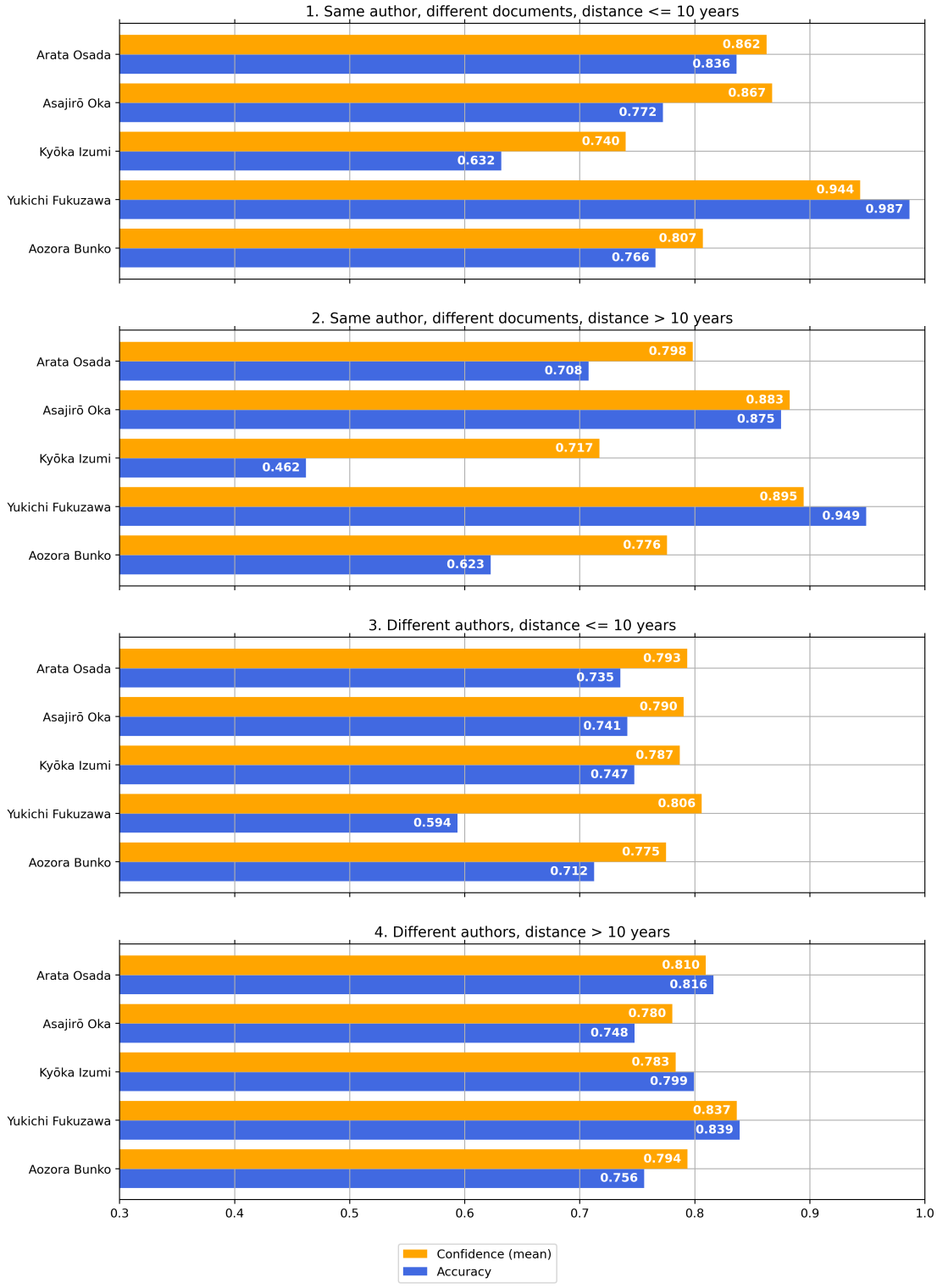

Figure 5.2: Results for the model trained on data without same-author samples.

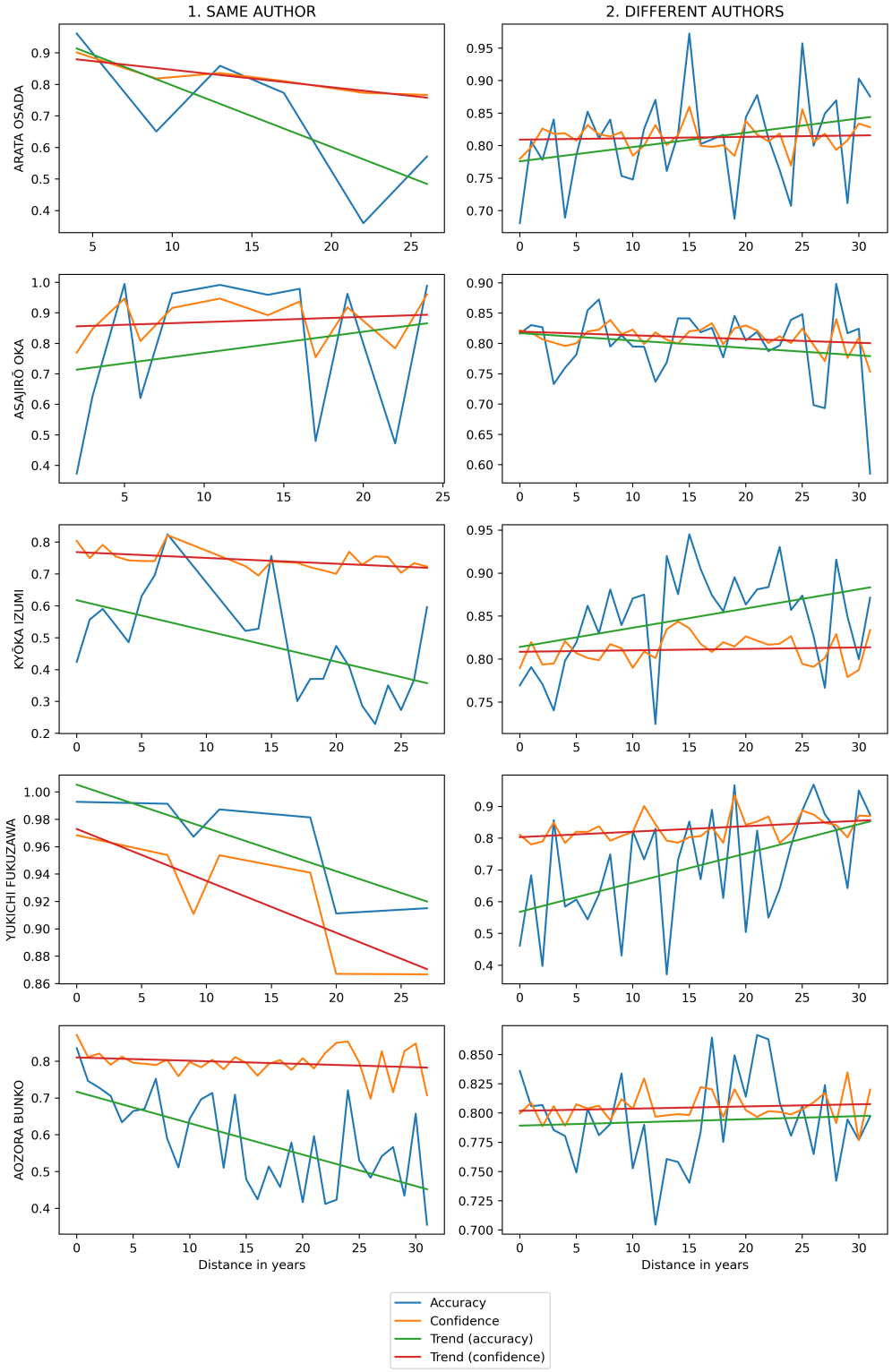

Figure 5.3: Relation between the distance in years between two documents and model's performance.

is founded. However the results of this experiment, particularly the extent to which the classification model's accuracy in detecting same authorship deteriorated when presented with texts written by the same author over a longer time span, show the need to take the aspect of time-related changes into consideration when performing authorship analysis.

That leaves us with the question of whether there actually exists any trait left consistent throughout one's writing, and if not – what factors (content? style?) specifically change in a person's writing over time, in what measure, how to quantify this change, how off can you be in predicting a writer's evolution and whether there exists a pattern of behavioral change observable through someone‘s writing. Tracing, evaluating and predicting such changes might be useful especially in the case of practical solutions in the field of forensics (terrorism countermeasures) or online bullying detection.

Another question is, from the point of view of the ethical evolution of an individual, what exactly constitutes a formative experience major enough to provoke an observable shift in opinions. In the case of the subject of my experiment, Arata Osada, the change in opinions can be traced back to his traumatic experience of the Second World War. In the case of different authors, who have not been subjected to major external historical stimuli, however, the question of whether a change is also noticeable due to a natural evolution of opinions juxtaposed with the passage of time, and how to quantify this evolution, remains open.

5.7 Conclusions

In this research I wanted to find out whether a model created for the task of same authorship detection would correctly attribute works from different points in time to the same author and with what accuracy. In particular, I was interested to see if the system experiences any additional difficulty in single authorship identification when presented with two texts by a person whose opinions and/or ethical values changed in the intervening period between writing them (e.g., in the case of the main object of this research, Arata Osada, a Japanese educator and specialist in the history of education, works written pre- and post-World War II) – which would mean that the impact of historical events on a person's ethical outlook and the content (books, articles, etc.) he or she produces, is significant enough that it can be quantified.

I conducted this research through performing a binary authorship analysis task using a text classifier based on a pre-trained transformer model, fine-tuned on randomly sampled pairs of text snippets from a repository of Japanese literary works (Aozora Library). The system was then tested on five different data sets, including a test set generated on the basis of four books authored by Arata Osada, (of which two were books written before the Second World War and another two, in the 1950s) and three comparison sets composed of works by three other Japanese authors: Yukichi Fukuzawa, Kyōka Izumi and Asajirō Oka. The task of the classifier was to identify whether there is one or many authors of the texts presented to it.

Upon performing the experiment, I found out that there is a strong negative correlation between the amount of time elapsed between the publication of two documents by a single person and the system's accuracy in detecting their common authorship, with the drop in performance being the highest for fiction books (i.e.,

novels), due to higher contextual diversity. In the group of non-fiction writers, the model had the lowest accuracy when presented with two fragments of Arata Osada's texts, one of them originating from before and one from after the Second World War. This and the observation that the classifier maintained a relatively high confidence in its incorrect judgements, indicate that a major shift in one's opinions as reflected in writing – although it might have less impact on the classification process than a change of topic – is often enough to convince the classification model that the authors are two different people, and by consequence that historical events such as a war can change the way an author expresses his opinions in writing beyond the point of recognition for a single authorship identification solution.

Chapter 6

Conclusions and Future Work

6.1 Summary and Future Directions

In this thesis, I presented the results of my research centered around implementing Culture, Religion and Time-Awareness to Machine Ethics Algorithms.
.

The main contributions of this research can be summarized as follows:

1. Experiment aimed at defining connotations of Buddhist terms with ethical criteria

2. Experiment for automatically determining whether changes in ethical education influence core moral values

3. Experiment on Authorship Analysis on the example of the works of Arata Osada

I began this thesis with a review of some of the related research in the area of AI and religion as well as AI and ethics. (Chapter 2).

In Chapter 3, I provide an overview of some of the previous cross-sectional studies in the field of Religion and Technology. After introducing the main resources and tools used in performing this research, I then report the outline and results of a preliminary experiment aimed at analyzing how much religious vocabulary, in particular Buddhist vocabulary taken from the largest online dictionary of Buddhist terms, is present in everyday social space of Japanese people, particularly, in Japanese blog entries appearing on a popular blog service (Ameba blogs).

While the general reaction to several expressions using Buddhist terms was as expected (for example the term *rinne*, 輪廻, the belief that all living things repeatedly pass through life and death, like a continually spinning wheel, was in majority met with emotions ranging from "resignation" to "suffering"), there were sometimes surprising twists in terms of social consequences. One such instance was the total of social consequences for the term *sōryo ni naru*, 僧侶になる, "to become a monk", which were unexpectedly indicating that this action was considered unethical at a staggering 66% (while the Buddhist worldview does treat becoming a monk as one of the most noble life objectives, as it gives more opportunities to attain awakening).

Also, concepts expressed in terms such as *gedō*, 外道 ("blasphemer", representative of a faith other than the Buddhist one), seemed to have lost their original meaning and instead became slang expressions, used to differentiate admirers of certain cultural phenomena of those critical of it as well as to distinguish between those doing things in a canonical way and those who don't.

In the future I plan to verify which dictionaries of Buddhist terminology make a distinction between terms still widely in use and those that have shifted meanings or became obsolete. I would also like to deepen my research through corroborating

the list of vocabulary obtained by filtering them through a list of Buddhist terms used in everyday modern Japanese. This would make it possible firstly to check how many terms are nowadays used regardless of their religious provenience and secondly to evaluate what percentage of vocabulary and in which categories (such as proverbs, art-related terms, etc.) have Buddhist origin. A bigger database of Buddhist terms as used in Japanese in the past and nowadays would also make it possible to track changes occurring in the structure of the Japanese vocabulary.

In Chapter 4, I described how to automatically determine whether changes in ethical education influenced core moral values in humans throughout the century. I analysed ethics as taught in Japan before the Second World War and today to verify how much the pre-Second World War moral attitudes have in common with those of contemporary Japanese, to what degree what is taught as ethics in school overlaps with the general population's understanding of ethics, as well as to verify whether a major reform of the guidelines for teaching the school subject of "ethics" at school after 1946 has changed the way common people approach core moral questions (such as those concerning the sacredness of human life).

I found out that past and contemporary Japanese share a similar moral outlook on certain basic moral concepts, although there is a discrepancy in the way in which they evaluate those concepts in terms of benefit to society and emotional consequences, which would indicate awareness in the society of the dissimilarity between the categories of ethics understood as rules provided by an external source and morals understood as an individual's own principles regarding proper and improper conduct.

Some phrases containing an ethical evaluation associated with a nationalistic trend in education, assessed positively in the pre-Second World War sample have

also disappeared from the scope of interest of post-war society, judging by their absence in a corpus of modern Japanese langauge. The findings of this study support suggestions proposed by others that the development of personal AI systems requires supplementation with moral reasoning, through indicating that while ethics is subject to evolution, due for instance to historical events or political agendas, there exists the need to balance them with morals, which at its core remain largely uninfluenced by external factors. Regardless, if (and when) personal AI systems obtain a singularity (a self-conscious state [30]), in order for the technology of personal artificial companions and life-long learning to respond to the private needs of a diversified range of users it must include an insight into culture-related and morality-related, pluralistic demands and needs of its potential users.

As a next step to deepen the analysis of this research, my plan is to perform the same process on two more samples, one related to work matters (relationships within one's workplace as well as with subordinates and superiors) and one related to patriotism (feelings towards one's home country), as those two areas of ethics are covered by both textbooks and we can expect a shift occurring in post-war Japan with the professional landscape dominated by big companies and the shift of the country's politics from militarism towards pacifism.

Furthermore, I would like to find out whether, as the results of my second experiment would suggest, the Japanese really lost interest in ethical matters related to concepts such as the motherland or the nation, through performing a survey analysing their approach, assessment and means to express patriotism.

Finally, in the course of a third experiment described in the chapter 5, and based on the findings of the two previous ones, I tried to answer a twofold question. I wanted to find out whether a model created for the task of same authorship

detection would correctly attribute works from different points in time to the same author and with what accuracy. In particular, I was interested to see if the system experiences any additional difficulty in single authorship identification when presented with two texts by a person whose opinions and/or ethical values changed in the intervening period between writing them (e.g., in the case of the main object of this research, Arata Osada, works written pre- and post-World War II) – which would mean that the impact of historical events on a person's ethical outlook and the content (books, articles, etc.) he or she produces, is significant enough that it can be quantified.

Upon performing the experiment, I found out that there is a strong negative correlation between the amount of time elapsed between the publication of two documents by a single person and the system's accuracy in detecting their common authorship, with the drop in performance being the highest for fiction books (i.e., novels), due to higher contextual diversity. This indicates that a major shift in one's opinions as reflected in writing – although it might have less impact on the classification process than a change of topic – is often enough to convince the classification model that the authors are two different people, and by consequence that historical events such as a war can change the way an author expresses his opinions in writing beyond the point of recognition for a single authorship identification solution.

I am planning to extend the analysis presented in this experiment by considering and comparing the effect of different aspects relevant to the process of authorship analysis, such as style (linguistic features) and content characteristics (semantics and topics).

As another direction of further research, I would like to analyze emotional

attitudes towards different ethically relevant concepts in the works of the same author (sentiment analysis) as well as define a list of text features that the model should take into consideration in order for it to be applied to other tasks such as authorship analysis in online harassment scenarios and broaden this set of features to make it adaptable for use with an artificial companion.

The most important findings of my research are that the development of personal AI systems requires supplementation with moral reasoning. Moreover, my whole research builds upon this idea and further suggests that AI systems need to be aware of ethics not as a constant, but as a function with a correction on historical and cultural changes in moral reasoning, influenced by several external factors, such as a religious outlook or lack of it as well as personal experiences or the history of a nation.

Bibliography

[1] A. Akbik, D. Blythe, and R. Vollgraf. Contextual String Embeddings for Sequence Labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

[2] W. R. Alger. *The theory of a personal devil*. 1861.

[3] M. Anderson and S. Anderson. *Machine Ethics*. Cambridge Univ. Press, 2011.

[4] M. Anderson and S. L. Anderson. Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine*, 28(4):15, Dec. 2007. DOI: 10.1609/aimag.v28i4.2065. URL: https://www.aaai.org/ojs/index.php/aimagazine/article/view/2065.

[5] S. Ariffin, A. A. Ismail, M. H. Yatim, and S. Sidek. An Assessment of Culturally Appropriate Design: A Malaysian University Context. *International Journal of Interactive Mobile Technologies (iJIM)*, 12(2):207–214, 2018. ISSN: 1865-7923. URL: https://online-journals.org/index.php/i-jim/article/view/8014.

[6] A. Ayer and B. Rogers. *Language, Truth and Logic*. Penguin Modern Classics. Penguin Books Limited, 2001. ISBN: 9780141911809. URL: https://books.google.co.jp/books?id=3M3ycPeuRNoC.

[7] H. Azarbonyad, M. Dehghani, M. Marx, and J. Kamps. Time-Aware Authorship Attribution for Short Text Streams. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 727–730, Santiago, Chile. Association for Computing Machinery, 2015. ISBN: 9781450336215. DOI: 10.1145/2766462.2767799. URL: https://doi.org/10.1145/2766462.2767799.

[8] D. Bagnall. Author Identification using Multi-headed Recurrent Neural Networks, June 2015.

[9] G. Barlas and E. Stamatatos. Cross-Domain Authorship Attribution Using Pre-trained Language Models. In I. Maglogiannis, L. Iliadis, and E. Pimenidis, editors, *Artificial Intelligence Applications and Innovations*, pages 255–266, Cham. Springer International Publishing, 2020. ISBN: 978-3-030-49161-1.

[10] J. Bentham. *An Introduction to the Principles of Morals and Legislation: The Collected Works of Jeremy Bentham*. Oxford University Press UK, 1996.

[11] A. Bovet and H. Makse. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, 10, Jan. 2019. DOI: 10.1038/s41467-018-07761-2.

[12] C. Chaski. Best Practices and Admissibility of Forensic Author Identification. *Journal of law and policy*, 21:5, 2013.

[13] N. Danaylov. Technology is the How, not the Why or What.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] J. Dower and W. N.

bibinitperiod Company. *Embracing Defeat: Japan in the Wake of World War II*. W.W. Norton & Company/New Press, 1999. ISBN: 9780393046861. URL: https://books.google.co.jp/books?id=gMElGZ93ZXAC.

[16] B. Duke. *The History of Modern Japanese Education: Constructing the National School System, 1872-1890*. Rutgers University Press, 2009. ISBN: 9780813544038. URL: https://books.google.co.jp/books?id=%5C_OOeA2H5aKIC.

[17] I. Frommholz, K. M, M. Potthast, Z. Ghasem, M. Shukla, and E. Short. *On Textual Analysis and Machine Learning for Cyberstalking Detection*. Bauhaus-Universität Weimar, 2017. URL: https://books.google.co.jp/books?id=A%5C_wXtAEACAAJ.

[18] R. M. Geraci. Spiritual robots: Religion and our scientific view of the natural world. *Theology and Science*, 4(3):229–246, 2006. DOI: 10.1080/14746700600952993. eprint: https://doi.org/10.1080/14746700600952993. URL: https://doi.org/10.1080/14746700600952993.

[19] S. Guthrie. *Faces in the Clouds: A New Theory of Religion*. Oxford University Press, 1995. ISBN: 9780195356809. URL: https://books.google.co.jp/books?id=dZNAQh6TuwIC.

[20] S. R. Hameroff, A. W. Kaszniak, and A. C. Scott. *Toward a Science of Consciousness: The First Tucson Discussions and Debates*. MIT Press, 1996.

[21] D. Holmes and J. Kardos. Who Was the Author? An Introduction to Stylometry. *CHANCE*, 16, Mar. 2003. DOI: 10.1080/09332480.2003.10554842.

[22] A. Husain. *The Sentient Machine: The Coming Age of Artificial Intelligence*. Scribner, 2017. ISBN: 9781501144677. URL: https://books.google.co.jp/books?id=hJg-DwAAQBAJ.

[23] M. Jakubíček, A. Kilgarriff, V. Kovář, P. Rychlý, and V. Suchomel. The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL 2013*, pages 125–127, Lancaster, 2013. URL: http://ucrel.lancs.ac.uk/cl2013/.

[24] P. Juola. Industrial Uses for Authorship Analysis. In 2015.

[25] I. Kant. *Grundlegung zur Metaphysik der Sitten [Groundwork of the Metaphysics of Morals]*. 1785.

[26] M. Kestemont, M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein, and M. Potthast. Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection. In *CLEF*, 2018.

[27] R. Komuda, R. Rzepka, and K. Araki. Aristotelian Approach and Shallow Search Settings for Fast Ethical Judgment. *International Journal of Computational Linguistics Research*, 4:14–22, Jan. 2013.

[28] R. Komuda, M. Ptaszynski, Y. Momouchi, R. Rzepka, and K. Araki. Machine Moral Development: Moral Reasoning Agent Based on Wisdom of Web-Crowd and Emotions. *International Journal of Computational Linguistics Research*, 1:155–161, Jan. 2010.

[29] M. Koppel and Y. Winter. Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1):178–187, 2014. DOI: 10.1002/asi.22954. eprint: https:

//asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.22954. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.22954.

[30] R. Kurzweil. *The Singularity is Near: When Humans Transcend Biology*. A Penguin Book: Science. Viking, 2005. ISBN: 9780670033843. URL: https://books.google.co.jp/books?id=88U6hdUi6D0C.

[31] G. Lample and A. Conneau. Cross-lingual Language Model Pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[32] K. MacDorman and S. Entezari. Individual differences predict sensitivity to the uncanny valley. *Interaction Studies*, 16:141–172, Sept. 2015. DOI: 10.1075/is.16.2.01mac.

[33] J. Maciejewski, M. Ptaszynski, and P. Dybala. Developing a Large-Scale Corpus for Natural Language Processing and Emotion Processing Research in Japanese. In *Proceedings of the International Workshop on Modern Science and Technology*, pages 192–195, 2010.

[34] D. Madigan, A. Genkin, D. Lewis, S. Argamon, D. Fradkin, L. Ye, and D. Consulting. Author Identification on the Large Scale, Jan. 2005.

[35] J. McGrath. *Religion and Science Fiction*. Wipf and Stock Publishers, 2011. ISBN: 9781621890249. URL: https://books.google.co.jp/books?id=EAVTAwAAQBAJ.

[36] T. C. Mendenhall. THE CHARACTERISTIC CURVES OF COMPOSITION. *Science*, ns-9(214S):237–246, 1887. ISSN: 0036-8075. DOI: 10.1126/science.ns-9.214S.237. eprint: https://science.sciencemag.org/content/ns-

9/214S/237.full.pdf. URL: https://science.sciencemag.org/content/ns-9/214S/237.

[37] MEXT. *Watashitachi no dōtoku – chugakkō [Our morality – junior high school]*. Monbukagakushō, 2014.

[38] M. L. Minsky. *The society of mind*. Simon and Schuster, 1986. ISBN: 9780671657130.

[39] A. Mohsen, N. El-Makky, and N. Ghanem. Author Identification Using Deep Learning. In pages 898–903, Dec. 2016. DOI: 10.1109/ICMLA.2016.0161.

[40] M. Mori, K. F. MacDorman, and N. Kageki. The Uncanny Valley [From the Field]. *IEEE Robotics Automation Magazine*, 19(2):98–100, June 2012. ISSN: 1558-223X. DOI: 10.1109/MRA.2012.2192811.

[41] N. Mori. *Shūshin kyōjuroku [Records from lectures in moral education]*. Chichisen-sho, 1989.

[42] F. Mosteller and D. L. Wallace. Inference in an Authorship Problem. *Journal of the American Statistical Association*, 58(302):275–309, 1963. ISSN: 01621459. URL: http://www.jstor.org/stable/2283270.

[43] A. Nakamura. *Kanjō hyōgen jiten [Dictionary of Emotive Expressions (in Japanese)]*. Tōkyōdō, 1993.

[44] J. Nieuwazny, K. Nowakowski, M. Ptaszynski, R. Rzepka, F. Masui, and K. Araki. Does change in ethical education influence core moral values? Towards culture-aware morality model. *Cognitive Systems Research*:in press, May 2020.

[45] S. Nirkhi, R. Dharaskar, and V. M. Thakare. Authorship Identification using Generalized Features and Analysis of Computational Method. *Transactions*

on Machine Learning and Artificial Intelligence, Apr. 2015. DOI: 10.14738/tmlai.32.1064.

[46] A. Osada. *Kyōiku kihonhō*. Shinhyoron, 1957.

[47] A. Osada. *Kyōiku shisōshi*. Iwanami Shoten, 1931.

[48] A. Osada. *Nihon no unmei to kyōiku*. Bokushoten, 1953.

[49] A. Osada. *Shinkyōiku no kōsō-Amerika no bunka-kyōiku wo hihan shite*. Fenikkusu Shoin, 1949.

[50] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

[51] R. W. Picard. *Affective Computing]*. MIT Press, 1997.

[52] N. Potha and E. Stamatatos. A Profile-Based Method for Authorship Verification. In A. Likas, K. Blekas, and D. Kalles, editors, *Artificial Intelligence: Methods and Applications*, pages 313–326, Cham. Springer International Publishing, 2014. ISBN: 978-3-319-07064-3.

[53] M. Ptaszynski, P. Dybala, R. Rzepka, K. Araki, and F. Masui. ML-Ask: Open Source Affect Analysis Software for Textual Input in Japanese. *Journal of Open Research Software*, 5:16, 2017. DOI: https://doi.org/10.5334/jors.149.

[54] M. Ptaszynski, P. Dybala, R. Rzepka, K. Araki, and Y. Momouchi. YACIS: A Five-Billion-Word Corpus of Japanese Blogs Fully Annotated with Syntactic and Affective Information. In *Proceedings of the AISB/IACAP 2012 Sympo-*

sium: Linguistic And Cognitive Approaches To Dialogue Agents (LaCATODA 2012), 2012.

[55] M. Ptaszynski, R. Rzepka, K. Araki, and Y. Momouchi. Automatically Annotating A Five-Billion-Word Corpus of Japanese Blogs for Sentiment and Affect Analysis. *Computer Speech & Language*, 28:38, Jan. 2014. DOI: 10.1016/j.csl.2013.04.010.

[56] C. Qian, T. He, and R. Zhang. Deep Learning based Authorship Identification. In 2017.

[57] J. F. Reeves. *Computational morality: a process model of belief conflict and resolution for story understanding. PhD thesis*. Computer Science Department, UCLA, 1991.

[58] A. Rexha, M. Kroll, H. Ziak, and R. Kern. Authorship identification of documents with high content similarity. *Scientometrics*, 115:223–237, Apr. 2018. DOI: 10.1007/s11192-018-2661-6.

[59] R. Rzepka and K. Araki. What Statistics Could Do for Ethics? The Idea of Common Sense Processing Based Safety Valve. In *AAAI Fall Symposium on Machine Ethics, Technical Report FS-05-06*, pages 85–87, Nov. 2005.

[60] R. Sakurai. Impacts of Recent Education Reforms in Japan: Voices from Junior High Schools in Japan. *Journal of international cooperation in education*, 18:55–65, 2016.

[61] U. Sapkota, S. Bethard, M. Montes, and T. Solorio. Not All Character N-grams Are Created Equal: A Study in Authorship Attribution. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–102,

Denver, Colorado. Association for Computational Linguistics, May 2015. DOI: 10.3115/v1/N15-1010. URL: https://www.aclweb.org/anthology/N15-1010.

[62] Y. Sari, M. Stevenson, and A. Vlachos. Topic or Style? Exploring the Most Useful Features for Authorship Attribution. In *COLING*, 2018.

[63] T. Shimizu. Author Identification of Japanese works using Doc2Vec and BERT, June 2020.

[64] E. Stamatatos. A Survey of Modern Authorship Attribution Methods. *JASIST*, 60:538–556, Mar. 2009. DOI: 10.1002/asi.21001.

[65] C. Stevenson. *Ethics and Language*. Donald F. Koch American Philosophy Collection. Yale University Press, 1944. URL: https://books.google.co.jp/books?id=xD9HpgkcpAgC.

[66] J. Taylor. *The Rule of Conscience; Or, Bishop Taylor's Ductor Dubitantium Abridged. by Richard Barcroft, ... in Two Volumes. ... of 2; Volume 2*. Creative Media Partners, LLC, 2018. ISBN: 9781379374862. URL: https://books.google.co.jp/books?id=ePqztwEACAAJ.

[67] W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right From Wrong*. Nov. 2008. ISBN: 978-0199737970. DOI: 10.1093/acprof:oso/9780195374049.001.0001.

[68] Y. Yan, W. Qi, Y. Gong, D. Liu, N. Duan, J. Chen, R. Zhang, and M. Zhou. ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training. *arXiv preprint arXiv:2001.04063*, 2020.

[69] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc., 2019.

[70] M. Zehfuss. *War and The Politics of Ethics*. OUP, editor. 2018.

[71] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization, 2019. arXiv: 1912.08777 [cs.CL].

[72] R. Zheng, J. Li, H.-c. Chen, and Z. Huang. A framework for authorship identification of Online messages: Writing-style features and classification techniques. *JASIST*, 57:378–393, Feb. 2006. DOI: 10.1002/asi.20316.