ETHI-CA2 2016:
ETHics In Corpus Collection, Annotation & Application

Workshop Programme

09:00 – Introduction by Workshop Chair
Laurence Devillers

09:10 – Keynote
Chairperson: Laurence Devillers
Edouard Geoffrois, *Interactive System Adaptation: Foreseen Impacts on the Organisation and Ethics of System Development*

09:50 – Talk 1
Chairperson: Björn Schuller
Teresa Scantamburlo and Marcello Pelillo, *Contextualizing Privacy in the Context of Data Science*

10:10 – Talk 2
Chairperson: Björn Schuller
Kevin Bretonnel Cohen, Karen Fort, Gilles Adda, Sophia Zhou and Dimeji Farri, *Ethical Issues in Corpus Linguistics And Annotation: Pay Per Hit Does Not Affect Effective Hourly Rate For Linguistic Resource Development On Amazon Mechanical Turk*

10:30 – Coffee and Poster session
Chairperson: Laurence Devillers

Jana Diesner and Chieh-Li Chin, *Gratis, Libre, or Something Else? Regulations and Misassumptions Related to Working with Publicly Available Text Data*

Wessel Reijers, Eva Vanmassenhove, David Lewis and Joss Moorkens, *On the Need for a Global Declaration of Ethical Principles for Experimentation with Personal Data*

Agnes Delaborde and Laurence Devillers, *Diffusion of Memory Footprints for an Ethical Human-Robot Interaction System*

Björn Schuller, Jean-Gabriel Ganascia and Laurence Devillers, *Multimodal Sentiment Analysis in the Wild: Ethical considerations on Data Collection, Annotation, and Exploitation*

Jocelynn Cu, Merlin Teodosia Suarez and Madelene Sta. Maria, *Subscribing to the Belmont Report: The Case of Creating Emotion Corpora*

Lucile Béchade, Agnes Delaborde, Guillaume Dubuisson Duplessis and Laurence Devillers, *Ethical Considerations and Feedback from Social Human-Robot Interaction with Elderly People*

11:10 – Talk 3
Chairperson: Joseph Mariani
Joss Moorkens, David Lewis, Wessel Reijers and Eva Vanmassenhove, *Language Resources and Translator Disempowerment*

11:30 – Talk 4
Chairperson: Joseph Mariani
Simone Hantke, Anton Batliner and Björn Schuller, *Ethics for Crowdsourced Corpus Collection, Data Annotation and its Application in the Web-based Game iHEARu-PLAY*

11:50 – Talk 5
Chairperson: Joseph Mariani
Agnes Delaborde, Noémie Enser, Alexandra Bensamoun and Laurence Devillers, *Liability Specification in Robotics: Ethical and Legal Transversal Regards*

12:10 – Panel and open discussion
Chairperson: Björn Schuller

12:30 – Conclusion
Laurence Devillers

Editors

Laurence Devillers	LIMSI-CNRS/Paris-Sorbonne University, France
Björn Schuller	Imperial College London, UK/University of Passau, Germany
Emily Mower Provost	University of Michigan, USA
Peter Robinson	University Cambridge, UK
Joseph Mariani	IMMI/LIMSI-CNRS/Paris-Saclay University, France
Agnes Delaborde	LIMSI-CNRS/CERDI/Paris-Saclay University, France

Workshop Organizers/Organizing Committee

Laurence Devillers	LIMSI-CNRS/Paris-Sorbonne University, France
Björn Schuller	Imperial College London, UK/University of Passau, Germany
Emily Mower Provost	University of Michigan, USA
Peter Robinson	University Cambridge, UK
Joseph Mariani	IMMI/LIMSI-CNRS/Paris-Saclay University, France
Agnes Delaborde	LIMSI-CNRS/CERDI/Paris-Saclay University, France

Workshop Programme Committee

Gilles Adda	LIMSI-CNRS, France
Jean-Yves Antoine	University of Tours, France
Nick Campbell	TCD, Ireland
Alain Couillault	GFII, France
Anna Esposito	UNINA, Italy
Karën Fort	Universite´ Paris-Sorbonne, France
Jean-Gabriel Ganascia	UPMC, France
Alexei Grinbaum	CEA, France
Hatice Gunes	Queen Mary University of London, UK
Dirk Heylen	University of Twente, Netherlands
Catherine Tessier	ONERA, France
Isabelle Trancoso	INESC, Portugal
Guillaume Dubuisson Duplessis	LIMSI-CNRS, France

Table of contents

Author Index

Preface/Introduction

Description

The focus of ETHI-CA2 spans ethical aspects around the entire processing pipeline from speech and language, as well as multimodal resource collection and annotation, to system development and application. In the recent time of ever-more collection "in the wild" of individual and personal multimodal and multi-sensorial "Big Data", crowd-sourced annotation by large groups of individuals with often unknown reliability and high subjectivity, and "deep" and autonomous learning with limited transparency of what is being learnt, and how applications such as in health or robotics depending on such data may behave, ethics have become more crucial than ever in the field of language and multimodal resources. This makes ethics a key concern of the LREC community. There is, however, a surprising if not shocking white spot in the landscape of workshops, special session, or journal special issues in this field, which ETHI-CA2 aims to fill in.

The goal is thus to connect individuals ranging across LREC's fields of interest such as human-machine and robot- and computer-mediated human-human interaction and communication, affective, behavioral, and social computing whose work touches on crucial ethical issues (e.g. privacy, traceability, explainability, evaluation, responsibility, etc.). According systems increasingly interact with and exploit data from humans of all ranges (e.g. children, adults, vulnerable populations) including non-verbal and verbal data occurring in a variety of real-life contexts (e.g. at home, the hospital, on the phone, in the car, classroom, or public transportation) and act as assistive and partially instructive technologies, companions, and/or commercial or even decision-making systems. Obviously, an immense responsibility lies at the different ends from data recording, labeling, and storage, to its pro-cessing and usage.

Motivation

Emerging interactive systems have changed the way we connect with our machines, modifying how we socialize, our reasoning capabilities, and our behavior. These areas inspire critical questions centering on the ethics, the goals, and the deployment of innovative products that can change our lives and society. Many current systems operate on private user data, including identifiable information, or data that provides insight into an individual's life routine. The workshop will provide discussions about user consent and the notion of informed data collection.

Cloud-based storage systems have grown in popularity as the scope of user-content and user-generated content has greatly increased in size. The workshop will provide discussions on best practices for data annotation and storage and evolving views on data ownership.

Systems have become increasingly capable of mimicking human behavior through research in affective computing. These systems have provided demonstrated utility, for interactions with vulnerable populations (e.g. the elderly, children with autism). The workshop will provide discussions on considerations for vulnerable populations.

The common mantra for assistive technology is, "augmenting human care, rather than replacing human care". It is critical that the community anticipates this shift and understands the implication of machine-in-the-loop diagnostic and assessment strategies.

Topics of interest

Topics include, but are not limited to:

- Ethics in recording of private content
- Ethics in multimodal, sensorial data collection
- Ethics in annotation (crowd-sourced) of private data Data storage/sharing/anonymization
- Transparency in Machine Learning

- Ethics in Affective, Behavioural, and Social Computing
- Responsibility in Educational Software and Serious Games Human-machine interaction for vulnerable populations
- Computer-mediated Human-Human Communication Responsibility in Decision-Support based on Data
- The role of assistive technology in health care

Summary of the call

The ETHI-CA2 2016 workshop is a crucially needed first edition in a planned for longer series. The goal of the workshop is to connect individuals ranging across LREC's fields of interest such as human-machine and robot- and computer-mediated human-human interaction and communication, affective, behavioural, and social computing whose work touches on crucial ethical issues (e.g. privacy, traceability, explainability, evaluation, responsibility, etc.). These areas inspire critical questions centering on the ethics, the goals, and the deployment of innovative products that can change our lives and consequently, society. It is critical that our notion of ethical principles evolves with the design of technology. As humans put increasing trust in systems, we must understand how best to protect privacy, explain what information the systems record, the implications of these recordings, what a system can learn about a user, what a third party could learn by gaining access to the data, changes in human behavior resulting from the presence of the system, and many other factors. It is important that technologists and ethicists maintain a conversation over the development and deployment lifecycles of the technology. The ambition of this workshop is to collect the main ethics, goals and societal impact questions of our community including experts in sociology, psychology, neuroscience or philosophy. At LREC 2016, the workshop shall encourage a broad range of its community's researchers to reflect about and exchange on ethical issues inherent in their research, providing an environment in which ethics co-evolve with technology.

Contextualizing Privacy in the Context of Data Science

Teresa Scantamburlo, Marcello Pelillo
DAIS - Università Ca' Foscari Venezia
via Torino 155, 30172 Mestre - Venezia
scantamburlo@dais.unive.it, pelillo@dais.unive.it

Abstract

Privacy is one of the most long-standing social issues associated to the development of information and communication technology and, over the years, from the diffusion of large databases to the rise of the World Wide Web and today's Internet of Things, just to name a few examples, the concerns have been intensified with consequences on the public discussion, design practices and policy making. Unfortunately the plethora of technical details on this topic has often discouraged people from tackling the problem of privacy and, hence, from being really active in the discussion of proposed approaches and solutions. In this paper we would like to provide fresh motivations to the inclusion of privacy in the overall pipeline of data processing, making this notion and its related issues somewhat more accessible to a non-expert audience. We will do that not by promoting new design standards or methodologies, which would add further technicalities, but by developing a critical perspective on the current approaches. In this way, we aim at providing data specialists with novel conceptual frameworks (e.g. the theory of conceptual integrity) to evaluate and better understand the place of privacy in their own work.

Keywords: Privacy, Data Science, Anonymity and Informed Consent, Contextual Integrity

1. Motivations

There is little doubt that the value of privacy has tremendously increased over the recent few years and the ascent of big data has clearly triggered this evolution. The ever-growing capacity to collect information about groups and individuals generating further knowledge from suitable analytics has made it clear that the concept of privacy cannot be circumvented or left to the expertise of lawyers or moral philosophers. Indeed, dealing with privacy is a complex activity that subsumes the participation of all the actors of the big data scenario, including the data scientists and the big data practitioners. Unfortunately the plethora of technical details on this topic has often discouraged people from tackling the problem of privacy and, hence, from being really active in the discussion of proposed approaches and solutions.

In this paper we would like to provide fresh motivations to the inclusion of privacy in the overall pipeline of data processing, from data collection to data analytics, making this notion and its related issues somewhat more accessible to a non-expert audience. Our aim is to challenge the idea according to which privacy is an additional attribute that can improve the overall assessment of data infrastructures (e.g., data generation, data collection, data storage, data analytics, etc.), but that has no intrinsic relevance to the work of engineers and computer scientists.

On the contrary, our assumption is that injecting some ideas about privacy, i.e. 1) what it means; 2) why it matters; 3) how it is treated and which critical aspects affect the current debate, into the expertise of those people could significantly improve the regulation of privacy issues and, more in general, the development of an ethical approach to data and data analytics. In particular, it would put data savvy professionals in the position to better understand the crucial role of privacy in the context of data science and, accordingly, to appreciate the profound social value of their own activities. In this way it could be easier to embed privacy within the development and the deployment of data-driven technologies supporting, from their own perspective, the creation of ethical guidelines and privacy policies.

2. Privacy and its theoretical challenges

Dealing with privacy is deeply problematic at least for two reasons. In the first place, there are difficulties in defining privacy at a conceptual level. Indeed, many have attempted to define what privacy is and why it is important. But the results are often controversial and the adequacy of proposed semantics is at the heart of passionate debates. In the second place, the reception of privacy has remarkably changed across society and, in fact, its value seems to be definitely nuanced in recent years. For example, some people think that nowadays respecting privacy has become almost impractical due to the inexorable growth of digital technologies and the huge number of activities based on it. While others argue that there is no threat to privacy if an individual is not engaged in illegal activities ("if someone has nothing to hide what is the problem with data disclosure?"), a position that is also known as the "nothing to hide" argument (Solove, 2007).

2.1. Some standard conceptions of privacy

Many a scholars have tried to conceptualize privacy around core concepts or by isolating a number of characterizing features. For example in (Solove, 2002) we find out that various attempts of defining privacy can be grouped under six general categories [1]:

[1] We briefly summarize the scheme proposed by Daniel Solove referring the interested reader to (Solove, 2002) for a full discussion. Note that these categories do not aim at providing a taxonomy of existing definitions. Rather, they represent an attempt "to track how scholars have chosen to theorize about privacy." (Solove, 2002, p. 1092). More recently Helen Nissenbaum organized some of the most prominent theoretical accounts around three fundamental dichotomies: "normative vs. descriptive accounts", "access-based vs. control-based accounts", "definitions based on the capacity to promote important values vs. definitions

1. *The right to be let alone.* This conception was formulated in 1980 by Samuel Warren and Louis Brandeis in their famous article *The Right to Privacy* (Warren and Brandeis, 1890), where privacy is fundamentally identified with the state of being inviolate and immune from any external assault. According to these authors, the value of privacy "is found not in the right to take the profits arising from publication, but in the peace of mind or the relief afforded by the ability to prevent any publication at all." (Warren and Brandeis, 1890, p. 200). Over the years, their position was synthesized by the expression "the right to be let alone", a way to highlight that privacy involves the respect to live one's life free from intrusion except when it is urged by community living.

2. *Limited access to the self.* Even if this vision could be considered similar to the first one, the limited-access conception of privacy is more articulated and in principle more extensive than the idea of being apart from others. One of the early formulation was given by Edwin Godkin who described privacy as the "right to decide how much knowledge of personal thought and feeling...private doings and affairs...the public at large shall have." (Godkin, 1890, 65). Another famous account was provided by Ruth Gavison's work, which is basically motivated by the idea of building a normative account upon a neutral, coherent description of privacy. According to her privacy can be measured in terms of "the degree of access others have to you through information, attention, and proximity."(Nissenbaum, 2010, p. 70).

3. *Secrecy.* The privacy-as-secrecy conception includes not only the interest of being let alone but also that of concealing private information. According to this view, any release of information is considered to be a reduction or a violation of privacy in particular when such a release determines a degradation of self-interest. With this conception, indeed, privacy is often viewed as a form of self-interested economic behavior since the concealment of information could be searched for one's own gain. Richard Posner, for instance, argued that when people strive for privacy protection they basically "want more power to conceal information about themselves that others might use to their disadvantage."(Posner, 1998, p. 271).

4. *Control of personal information.* Privacy as the control over personal information is probably the one of the most influential theories of privacy. Indeed, as Helen Nissenbaum suggested, ranging from law to policy many conceptions of privacy incorporate the notion of control as a key aspect. This view can be traced back to Alan Westin's celebrated book *Privacy and Freedom*, where the author acknowledged that privacy is "the claim of individuals, groups, or institutions to determined for themselves when, how, and to what extent information about them is communicated to others." (Westin, 1967, p. 7). Along these lines, Charles Fried argued that "Privacy is not simply an absence of information about us in the minds of others; rather it is the control we have over information about ourselves."(Fried, 1968, p. 482).

5. *Personhood.* Another theoretical account views privacy as a way to protect personhood, a term coined by Paul Freund to define "those attributes of an individual which are irreducible in his selfhood."(Freund, 1975). The main aim of this theory is to protect the integrity of personality and its employment is often conceived as a support for other frameworks. Relating privacy to the idea of personhood is particularly useful to emphasize why privacy is important and what aspects of the self privacy should protect. In these respects various scholars pointed out that privacy is grounded on important moral value, such as personal dignity, autonomy, self-determination, so that any form of intrusion or surveillance could represent a limitation of individual self-expression.

6. *Intimacy.* With respect to moral personhood, the theorists of privacy as intimacy seek to shift attention from individual self-creation to personal relationships. As Solove suggested by "focusing on the relationship-oriented value of privacy, the theory of privacy as intimacy attempts to define what aspects of life we should be able to restrict access to, or what information we should be able to control or keep secret."(Solove, 2002, p. 1121).

All these theoretical accounts could highlight several peculiar traits of privacy (e.g., the need to regulate the access to information by others and to control the flow of information, or its reference to individual autonomy and intimate relationship, etc.) and, considered as a whole, they indeed reflect the multifaceted nature of such a complex notion.
However, according to Solove they all suffer from being either too restrictive or excessively vague and, as a result, they turned out to be inadequate to solve concrete problems. For example the attempt to define privacy in terms of intimacy has been judged too narrow for not all private information or decisions can be considered intimate at any time and much of this evaluation depends on the context we are referring to (e.g., speaking about personal religious or philosophical belief would not be probably considered intimate information in caring communities or therapeutic groups). By contrast, the position that views privacy as "the right to be let alone" has been considered too ambiguous and replete with several unsolved questions (e.g., "what does it mean concretely "being let alone"?", "under which conditions is it retained to be satisfied?" and so on).

2.2. A pragmatic account of privacy

In order to overcome the problem of finding the "essence" of privacy, Solove advocated a more pluralistic understanding, similar in spirit to Wittgenstein's family resemblance, where, instead of looking for a common denominator with a universal value, the main aim is to focus on the specific practices which require privacy protection. Based on this

based on the capacity to protect a specific, private realm" (Nissenbaum, 2010)

pragmatic view, Solove ends up with a taxonomy which examines the various problems and harms that may constitute a breach of privacy (Solove, 2006). According to this classification there are four groups of activities, spanning from information collection to dissemination and invasion (see Table 2.2. for a scheme of this taxonomy), but in our discussion we will be focused on two of them: information collection and information processing.

A first group of activities that might violate privacy includes the processes of *information collection*, such as surveillance and interrogation. Note that, in the case of interrogation, that is the "the pressuring of individuals to divulge information", the harms against privacy could be less apparent than what one would expect. Indeed, with the World Wide Web and the other technologies composing the attractive paradigm of the Internet of Things (e.g., Radio-Frequency IDentification, Near Field Communication, Bluetooth Low Energy, etc.) it became increasingly easy to leave digital footprints about our own behavior on the Internet.

At first glance, these footprints could be considered limited in scope, as regarding distinct spheres of life (purchases, health, particular hobbies, etc.). But the trend of today's technology is to combine all these partial perspectives to create, thanks to appropriate data analysis techniques, a more powerful pictures of the world, including people, objects, institutions and territories. Interestingly, Solove observed that potential distortions and manipulations of data may occur even during information collection, that is at the early phases of the information cycle, before any data-processing mechanism really starts. This possibility is associated to the extraordinary power of interrogation: That of controlling what information can be elicited. Indeed, as Solved put it "a skillful interrogator can orchestrate a dialog that creates impressions and inferences that she wants to elicit."(Solove, 2006, 501)

A second important group of activities regards *information processing* which includes the entire spectrum of actions concerning the use, storage and manipulation of collected data. Among these Solove places "aggregation", a number of practices involving the attempt of gathering information, linking data and, in this way, the effort of learning further insights that isolated information would not reveal. In the present data deluge, aggregation has become a strategic component of information technology systems, and much of its development has been supported by the advances of machine learning and data mining techniques. With this methods, machines can extract novel information (correspondences, similarities, patterns, etc.) from huge data-sets and create models that can be used to make predictions and evaluate experimental results.

Associated with the growth and the diffusion of machine learning and data mining techniques there is also the activities concerning identification and, specifically, the ability of connecting data to particular human beings. Identification, indeed, is intrinsically related to the work of machine learning, where the aim is to associate certain characteristics to individual, either objects or humans. This may results beneficial in many cases, e.g. when this helps to prevent crimes or social harms, but it could be discriminatory when correlating vast amounts of data (including sensitive information) algorithms produced unfair classifications or decisions (e.g. about health care, employment, housing) based on analyst's prior prejudices or the biases within society (Barocas and Selbst, 2014).

3. The current debate

The aforementioned difficulties at a conceptual level have had various implications on the current debate. On the one hand, they emphasized a general resigned pessimism which in turn produced very radical conclusions, like Robert Post's declaration: "privacy is a value so complex, so entangled in competing and contradictory dimensions, so engorged with various and distinct meanings that I sometimes despair whether it can be usefully addressed at all." (Post, 2001).

On the other hand, the increasing awareness of privacy complexity has drown public attention to the urgency of practical provisions enabling companies, institutions and organizations to balance the enormous advantages stemming from data-centric technologies with the need for protecting personal information. From a practical point of view, this pressure has resulted in a proliferation of rules, guidelines and safe practices which ought to give an outlet for the ethical worries affecting many technology-based activities. In this way, privacy protection would be guaranteed by the incorporation of such practices within the wide rage of activities that might imply the disclosure of personal information (e.g., those included in Solove's taxonomy).

In the following subsection we will briefly introduce the basic procedures provided to protect privacy and in particular the tool of notice and consent, one of the most used technical solutions for privacy protection. We will refer to these practices by the label "procedural approach" not to confine these methods within specific boundaries but just to stress the persistence of an operational attitude in privacy policies. Note that a procedural mentality is somehow reflected also by the more recent attempts of incorporating privacy within algorithms and within the design process. This is the case, for instance, of statistical frameworks (Karr and Reiter, 2014), the model of differential privacy (Dwork, 2006) and other formal methods for enforcing privacy policy (e.g. see (Datta, 2014)).

3.1. The procedural approach

At present, one of the most common way to protect privacy is to develop appropriate procedures and specific protocols that fulfill at least tow tasks: to limit external access to personal information and to assure people's right to control the disclosure of personal information (i.e., when, how, and to what extent information about them can be communicated to others). With respect to the theoretical accounts examined in the previous section, this approach makes reference to some specific ideas of privacy and in particular to those based on the access and the control of information. Note that a strong assumption of the control-based or access-based solutions regards the role of the individual, who has to self-menage the the sphere of private, sensitive data.

DOMAIN	PRIVACY BREACH	DESCRIPTION
Information collection	Surveillance	Whatching, listening to, or recording of an individual's activities
	Interrogation	Various forms of questioning or probing for information
Information-processing	Aggregation	The combination of various pieces of data about a person
	Identification	Linking information to particular individuals
	Insecurity	Carelessness in protecting stored information from leaks and improper access
	Secondary use	Information collected for one purpose used for a different purpose without the data subject's consent
	Exclusion	Failure to allow the data subject to know about the data that others have about them and participate in its handling and use, including being barred from being able to access and correct errors in that data
Information dissemination	Breach of confidentiality	Breaking a promise to keep a person's information confidential
	Disclosure	Revelation of information about a person that impacts the way others judge their character
	Exposure	Revealing another's nudity, grief or bodily functions
	Increased accessibility	Amplifying the accessibility of informatics
	Blackmail	Threat to disclose personal information
	Appropriation	The use of the data subject's identity to serve the aims and interests of another
	Distortion	Dissemination of false or misleading information about individuals
Invasion	Intrusion	Invasive acts that disturb one's tranquillity or solitude
	Decisional inference	Incursion into the data subject's decisions regarding their private affairs

Table 1: Solove's Taxonomy. Source (Kitchin, 2014)

Now, the realm of personal data, also known as "sensitive information", embraces a wide rage of information and can include: ethnicity, gender, sexual orientation, age, religious beliefs, political opinion, membership, physical or mental health details, marital status, criminal records and so on. The list of sensitive data is complemented also by other types of information that can relate to individuals or their activity as consumers, employees, clients, students, patients, etc. Among these data there is also any information that can, more or less directly, identify a person. This list of data is codified in the so-called personally identifiable information (PII) and includes: contact information (e.g., postal address, e-mail address), birth date, social security numbers, credit or debit card numbers, driver's license number. But note that other sensitive data are represented by IP address, ID mobile number, cookies, which in turn may reveal information bout personal preference and activities.

Note that the procedural approach is well reflected in the Fair information Practice Principles (notice, choice, consent, security, integrity access and accountability), a central pillar for much of today's privacy regulation (Solove, 2013). The main idea underlying these principles is "to provide people with control over their personal data, and through this control people can decide for themselves how to weigh the costs and benefits of the collection, use, or disclosure of their information." (Solove, 2013, p. 1880). In general a solution stemming from this strategy is to limit the disclosure of personally identifiable information (PII) by anonymization techniques and the well-known informed consent (Kitchin, 2014).

Anonymization is the procedure which removes from collected data any details that could identify a person (i.e. personally identifiable information) or which makes little changes in order to avoid the identification of individuals. In the past few years various techniques have been developed in order to de-identify people and a vast literature has grown up around this topic (see, e.g. (Weber and Heinrich, 2012)). A possible solution could be data masking, that is the replacement of original (sensitive) data with a special characters (e.g., a sequence of "x") or data obfuscation which consists in substituting specific data with others preserving the format and the type. Many camps of real life solve privacy issues thanks to anonymization techniques and a lot of sensitive data flows abundantly in view of the fact they are guaranteed by anonymity.

The other technique frequently used to deal with privacy problems is informed consent, the process which regulates the disclosure of personal information to others and their use for some specific purposes (experimental analysis, surveys). Informed consent is a widespread practice which has found application in many different environment from health care to business activities.

Unfortunately anonymity and informed consent present various limitations even at a practical level. Indeed, as for anonymity "even when individuals are not 'identifiable', they may still be 'reachable', may still comprehensibly represented in records that detail their attributes and activities, and may be subject to consequential inferences and predictions taken on that basis."(J. Lane and Nissenbaum, 2014, p. 45). While, with respect to consent it has been suggested that ordinary experiences are in sharp contrast

with the ideal of privacy regulation as self-management. Indeed, as some social analyses pointed out, "(1) people do not read privacy policies; (2) if people read them, they do not understand them; (3) if people read and understand them, they often lack enough background knowledge to make an informed choice; and (4) if people read them, understand them, and can make an informed choice, their choice might be skewed by various decision-making difficulties."(Solove, 2013, p. 1888).
But most importantly, anonymity and consent has somehow inhibited an in-depth appreciation of privacy and its genuine interaction with other values (e.g., fairness, autonomy, justice and so on). Indeed, the proliferation of such mechanisms has somehow reinforced the idea that privacy is a practical matter, a problem that can be added to pile of technical details concerning data collection and analytics. But, in this way, we loose the source of privacy concerns: the reasons why privacy matters, which specific values and goals it serves, the moral conflicts underlying determinate information flow and, finally, the motivations to prefer one solution to another. With these worries in mind, we will sketch a different perspective in the following sections so as to shed light on these more fundamental aspects.

4. The contextual approach to privacy

An alternative way to approach privacy, from both a theoretical and a practical point of view, has been provided by Nissenbaum's framework of contextual integrity (Nissenbaum, 2010). This theory differs from the standard articulation of privacy at least for two main reasons. First, it does not root privacy in one, single characteristic, like secrecy or intimacy. Secondly it prefers to look at privacy not as a right to control or to limit access to information, but "as a right to appropriate flow of personal information."(Nissenbaum, 2010, p. 127). Hence, the theory of contextual integrity tends to shift the emphasis from restricting the flow of information to ensuring that it flows appropriately.
In particular, as for on-line activities, the theory of contextual integrity contends that respecting privacy requires a fully integrated approach and, thus, it tends to not separate privacy from the rest of social and moral values. Its motivating assumptions are:

1. that on-line activity is inextricably tied to the social life, i.e. it is not "a distinctive venue, sphere, place, or space defined by the technological infrastructures and protocols of the Net" (Nissenbaum, 2011, p. 38) ;

2. that on-line activity reflects the variety of off-line experience, i.e. "it is radically heterogeneous, comprising multiple social contexts, not just one." (Nissenbaum, 2011, p. 38).

As a consequence, the norms regulating on-line experience are context-sensitive and profoundly influenced by the social sphere in which such experiences take place (e.g., health care, education, employment). Indeed, within each of these contexts there exist, either implicitly or explicitly, diverse social norms (e.g., concerning roles, behaviors or expectations) which shape and limit human practices.

At the heart of this framework there are the idea of social context and that of informational norms. Basically, the theory relies on the robust intuition according to which individuals do not act in isolation but in a plurality of social contexts (education, health-care, family life, commercial marketplace, work life, etc.), structured on various social norms, habits and values. This means that even the norms which govern the exchange and the flow of information (e.g., transmission, distribution, dissemination, etc.) cannot be fully understood out of context. On the contrary, they capture essential aspects of the social settings in which information flows including social roles, social structures, etc. More in general, context-relative informational norms may have two functions: They "express entrenched expectations governing the flows of personal information, but they are also a key vehicle for elaborating the prescriptive (or normative) component of the framework of contextual integrity."(Nissenbaum, 2010, p. 129).
To articulate in more details the description of such norms, the framework of contextual integrity identifies a few parameters (Nissenbaum, 2010), such as:

- the context (the general conditions of application);
- the actors (subject, sender, recipient);
- the attributes (types of information)
- and the transmission principles (constraints under which information flows).

The main idea is that context-relative information norms govern what type and how much personal information is relevant and appropriate to be shared with others according to the considered social setting. Thus, following the theory of contextual integrity, for example, we would conclude that it makes sense to share with the physician details about physical conditions but not about financial investment or salary. Interestingly, as other examples would make it clear, the main contribution of this framework is to shed light on the reasons behind privacy breaches (e.g., highlighting the expectations, the social roles, norms and the specific values at stake) and, in a sense, to avoid the dichotomy between privacy complexity and privacy management.
Moreover to reinforce the normative value of context-relative informational norms, Nissenbaum has developed a number of decision heuristic that could offer farther support to the evaluation of concrete cases (Nissenbaum, 2010, p. 182). Her guidelines include:

1. Describe the new practice in terms of information flows.

2. Identify the prevailing context...and identify potential impacts from contexts nested in it.

3. Identify information subjects, senders, recipients.

4. Identify transmission principles.

5. Locate applicable entrenched informational norms and identify significant points of departure.

6. Prima facie assessment...A breach of information norms yields a prima facie judgment that contextual integrity has been violated because presumption favors the entrenched practice.

7. Evaluation I: Consider moral and political factors affected by the practice in question...

8. Evaluation II: Ask how the system or practices directly impinge on values, goals, and ends of the context...

9. On the basis of these findings, contextual integrity recommends in favor of or against systems or practices under study...

As compared with the standard definitional approach, the theory of contextual integrity, assuming that informational norms, like social norms, evolve and may undergo various cultural and societal alterations, it prefers to provide some criteria to distinguish which information practices could be morally problematic rather than prescribing fixed scheme to control information access and distribution. Therefore, the theory of contextual integrity does not provides any rules to discriminate a priori what is public from what is private, but, on the contrary, it offers a conceptual framework that model privacy with respect to the social and ethical background of the information flow.
Until now, Contextual Integrity has been introduced to deal with privacy issues in on-line activities, such as "blog sphere."(Grodzinsky and Tavani, 2010). However its main building blocks could be inspiring even for the study of machine learning at large, not only with respect to the protection of personal information but also to the communication and the application of predictive analytics.

5. Concluding remarks

In our sketch of privacy we have pointed out some intrinsic difficulties concerning the conceptualization of such a notion. This has been stressed by many a scholars to the point that "even the most strenuous advocate of a right to privacy must confess that there are serious problems of defining the essence and scope of this right." (Beaney, 1966, p. 255). In general, most of the theoretical efforts moved in the direction of an essential definition thereby restricting the notion of privacy to a valuable, but limited perspective (such as that of control or secrecy). The limitations of an all-encompassing definition are visible also at a practical level, when we have to specify what data should be considered as "personal" or "sensitive" and under which circumstances the flow of information has to be limited or controlled.
An alternative way to characterize privacy is suggested by Solove's pragmatic view. The latter tries to capture the notion of privacy by exploring the technological activities which could raise privacy concerns. Nevertheless, this approach does exclude the possibility to highlight some characterizing features but it overcomes the obstacle of a unique, essential definition. Moreover it offers the opportunity to extend the conception of privacy with possible new traits that are now unpredictable but that could emerge from future technological developments.
As for privacy polices, the dominant approach to addressing privacy issues has been represented by anonymity and informed consent, a solution which has generally preferred the regime of "take it or leave it" (Nissenbaum, 2011, p. 35). These solutions has emphasized the individual character of privacy which in the end becomes an expression of a self-interest.
A completely different approach is given by Nissenbaum's theory of contextual integrity, whose peculiar merit is to offer, instead of abstract norms, a critical perspective on the practices and the technologies affecting the flow of personal information. Interestingly, as we noticed above, such a theory tries to articulate the foundation of privacy policy and regulation by answering "questions not only of the form: *what* policies [...], *what* technical standards and design features, but *why* these." (Nissenbaum, 2010, p. 7). Moreover, this theory offers a set of guidelines that can be a further resource for the design and the evaluation of big data projects, such as, "describing the practice in terms of information flows", "identifying the prevailing context and potential impacts from contexts nested in it" and "identifying information subjects, senders, recipients" (Nissenbaum, 2011).
Our outline of privacy issues, as suggested before, is not intended to provide a specific ready-to-use procedure, but to discuss privacy in a critical way so as to uncover the social and ethical values behind it and give further conceptual tools to data scientists.

6. Bibliography

Barocas, S. and Selbst, A. (2014). Big data's disparate impact. *SSRN eLibrary*.

Beaney, W. (1966). The right to privacy and american law. *LAW and CONTEMP. PROBS.*, 31:253–271.

Datta, A., (2014). *Privacy through Accountability: A Computer Science Perspective*, pages 43–49. Springer International Publishing.

Dwork, C., (2006). *Differential Privacy*, pages 1–12. Springer Berlin Heidelberg.

Freund, P. (1975). Address to the american law institute. In *Proceedings of the 52nd Annual Meeting of The American Law Institute*, pages 42–43, May.

Fried, C. (1968). Privacy. *Yale Law Journal*, 77:475–493,.

Godkin, E. (1890). The rights of the citizen: Iv-to his own reputation. *SCRIBNER'S MAGAZINE*, July:58–67,.

Grodzinsky, F. and Tavani, H. (2010). "contextual integrity" model of privacy to personal blogs in the blogsphere. *nternational Journal of Internet Research Ethics*, 3:38–47.

J. Lane, V. Stodden, S. B. and Nissenbaum, H. (2014). *Privacy, Big Data, and the Public Good. Frameworks for Engagement*. Cambridge University Press.

Karr, A. and Reiter, J., (2014). *Using Statistics to Protect Privacy*, pages 276–295. Cambridge University Press.

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Their Infrastructures and Their Consequences*. Sage.

Nissenbaum, H. (2010). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus*, 140(4):32–48.

Posner, R. (1998). *Economic Analysis of Law*. Aspen, 5th edition.

Post, R. (2001). Three concepts of privacy. *GEORGETOWN LAW JOURNAL*, 89:2087–2098.

Solove, D. (2002). Conceptualizing privacy. *California Law Review*, 90 (4):1087–1155.

Solove, D. (2006). A taxonomy of privacy. *University of Pennsylvania Law Review*, 154(3):477–560.

Solove, D. (2007). I've got nothing to hide and other misunderstanding of privacy. *San Diego Law Review*, 44:745–757.

Solove, D. (2013). Privacy management and the consent dilemma. *Harvard Law Review*, 126:1880–903.

Warren, S. and Brandeis, L. D. (1890). The right to privacy. *Harvard Law Review*, 4:193–220,.

Weber, R. and Heinrich, U. (2012). *Anonymization*. Springer.

Westin, A. (1967). *Privacy and Freedom*. Atheneum.

Ethical Issues in Corpus Linguistics And Annotation:
Pay Per Hit Does Not Affect Effective Hourly Rate For Linguistic Resource Development On Amazon Mechanical Turk

K. Bretonnel Cohen[1], Karën Fort[2], Gilles Adda[3], Sophia Zhou[4], and Dimeji Farri[4]
[1] Biomedical Text Mining Group, Computational Bioscience Program, U. Colorado School of Medicine
[2] Equipe Sens Texte Informatique Histoire, Université Paris-Sorbonne
[3] Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
[4] Philips Research North America

Abstract

Ethical issues reported with paid crowdsourcing include unfairly low wages. It is assumed that such issues are under the control of the task requester. Can one control the amount that a worker earns by controlling the amount that one pays? 412 linguistic data development tasks were submitted to Amazon Mechanical Turk. The pay per HIT was manipulated through a range of values. We examined the relationship between the pay that is offered per HIT and the effective pay rate. There is no such relationship. Paying more per HIT does not cause workers to earn more: the higher the pay per HIT, the more time workers spend on them ($R = 0.92$). So, the effective hourly rate stays roughly the same. The finding has clear implications for language resource builders who want to behave ethically: other means must be found in order to compensate workers fairly. *The findings of this paper should not be taken as an endorsement of unfairly low pay rates for crowdsourcing workers. Rather, the intention is to point out that additional measures, such as pre-calculating and communicating to the workers an average hourly, rather than per-task, rate must be found in order to ensure an ethical rate of pay.*

Keywords: ethics, corpus linguistics, corpus annotation, Amazon Mechanical Turk, crowdsourcing

1. Introduction

Crowdsourcing has become a popular way to create data for research, in particular in natural language processing (NLP). There are a variety of approaches to crowdsourcing and many crowdsourcing taxonomies, many of which are presented in (Geiger et al., 2011). One way to distinguish between these many approaches is to consider (i) the remuneration of the participants and (ii) the transparency of the task (that is, whether or not it is obvious to the participants). This small set of features allows one to distinguish between three major types of crowdsourcing: (i) volunteer and transparent, as in the case of vested volunteers who have a personal commitment to the intended use of the data (Cohen et al., 2015); (ii) volunteer and not transparent, as in the case of games with a purpose (GWAPs), which offer an entertaining experience to the participants; and (iii) remunerated and transparent crowdsourcing, i.e. microworking. The latter is typically done via dedicated platforms such as Amazon Mechanical Turk or CrowdFlower, and raises a number of ethical issues. Some of these have been addressed in various publications, including (Fort et al., 2011).

Analysts have identified a number of ethical issues with paid crowdsourcing (Adda et al., 2013). Unfairly low wages (Ross et al., 2009; Chilton et al., 2010) are one such problem. As a significant proportion of the workers use MTurk as their primary source of income, or to make basic ends meet (Ross et al., 2010; Ipeirotis, 2010; Fort et al., 2011), this becomes an ethical issue. Those very low wages are partly induced by the pay per task model, because the worker is not aware of the hourly rate before choosing the task (Callison-Burch, 2014). Another frequently mentioned problem (Silberman et al., 2010) is the fact that requesters sometimes pay late, or even not at all.

It is widely assumed that these issues are under the control of the purchaser of crowdsourcing services. The work reported here investigates a number of assumptions about these issues and about the extent of purchaser control over them. In particular, we wondered: suppose that a purchaser of annotation services through a crowdsourcing site wants to ensure that they pay an ethical wage. Can one control the amount that a worker earns by controlling the amount that one pays? It seems obvious that one should be able to, but early experiences suggested that this might not, in fact, be the case.

The methodology was as follows. In the course of our normal work on preparing linguistic resources for use in developing and testing natural language processing applications, a variety of types of tasks were submitted to Amazon Mechanical Turk. The pay per HIT (Human Intelligence Task, the basic unit of work performance on the Amazon Mechanical Turk web site) was manipulated through a range of values (never below the typical payment for a task type). The total data set contains 412 data points.

The Amazon Mechanical Turk interface provides a number of data points upon completion of a task. These include:

- Pay per HIT: this is the amount offered per HIT by the person who "requests" (in Amazon Mechanical Turk parlance) that the work be done.
- Average time per assignment: this is the average amount of time spent by a worker on a HIT.
- Effective hourly rate: this is the extrapolated amount earned per hour by a typical worker for doing the task.
- Agreement: for classification tasks, this is the agreement between workers.
- Total number of HITs completed: this is the number of HITs done at the indicated pay per HIT, effective

hourly rate, etc.

Reasonable expectations are that all other things being equal:

- Effective hourly rate should correlate positively with pay per hit.
- Average time per assignment should correlate positively with pay per hit.
- Effective hourly rate should correlate negatively with average time per assignment.
- Agreement should correlate positively with pay per hit.
- Agreement should correlate negatively with effective hourly rate.

The reader may disagree with the authors' expectations about these relationships, but the data presented here allows the testing of discordant expectations, as well. That is, whether the reader agrees with the author that effective hourly rate should correlate positively with pay per HIT, or thinks that it should correlate negatively, or doesn't think that there should be any correlation at all, the data allows testing any of those hypotheses.

1.1. Tasks

Data from a variety of task types is analyzed here. Tasks were not created specifically for this paper—these were tasks that we carried out in the course of our normal research work. The task types discussed here are:

- Information extraction: relation annotation (1 set of tasks)
- Recognizing textual entailment: language generation (3 separate sets of tasks)
- Recognizing textual entailment: classification (1 set of tasks)
- Paraphrase relations: classification (3 separate sets of tasks)

The number of workers participating in a task is variable from one set of tasks to another, as is the number of subtasks (e.g. classifying a single pair of sentences versus writing three separate sentences) and the total number of completed HITs per task.

2. Results

The data consists of 412 completed HITs. There was no attempt to balance across the various pay levels or task types—the tasks were requested in the course of the authors' normal work. Table 1 gives descriptive statistics on the number of HITs completed at each level of pay per HIT. Table 2 gives the number of HITs completed for each task type.

Figure 1 shows the effective hourly rate for the various tasks as a function of the pay per HIT. It is clear from the figure that there is no relationship between the pay that is offered and the amount that is earned. Since there is no linear relationship, we do not calculate a correlation. The data show that we cannot cause workers to earn higher wages by paying more per HIT. Regardless of whether we pay $0.05 per HIT or five times that much, the effective hourly rate hovers around the median of US $2.25. It is worth noting that the one set of tasks that shows an apparent effective hourly rate of US $12.50 per hour had only 12 completed tasks, and therefore the sample size is much smaller than for the other sets of tasks.

Table 1: Number of HITs completed at each level of pay per HIT. The 8 sets of tasks comprised 412 individual hits.

Pay per HIT	number of HITs completed
US $ 0.05	110
US $0.10	173
US $0.25	129
Total HITs completed	412

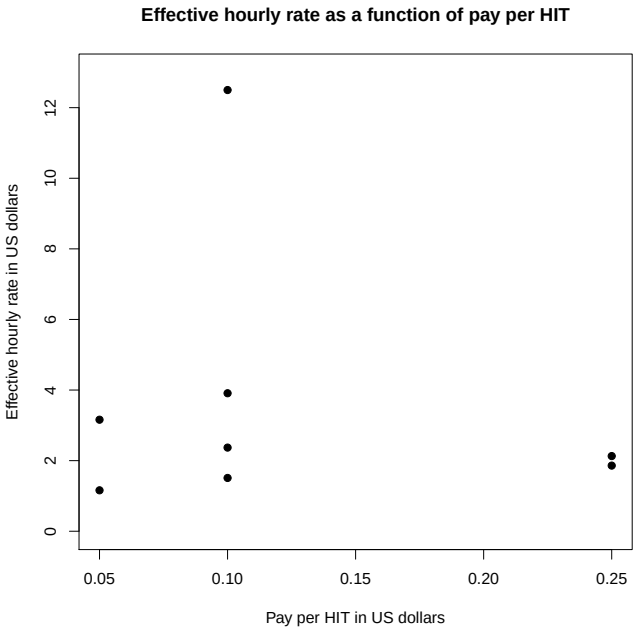

Figure 1: There is no relationship between the pay per HIT and the effective hourly rate earned by workers. Each data point on the graph represents a separate set of tasks. 412 HITs were completed in the course of these 8 sets of tasks.

Figure 2 shows the agreement for the various classification tasks as a function of the pay per HIT. Examining the inter-rater agreement on the five classification tasks as a function of pay per HIT, it does not appear that there is a relationship between the pay that is offered and the agreement that is achieved. Since there is no linear relationship, we do not calculate a correlation. The data show that we cannot get better agreement by paying more per HIT. The agreement that is achieved at a pay per HIT of $0.25 is not necessarily

Table 2: Number of HITs completed for each task type. The 8 sets of tasks comprised 412 individual hits.

Task	number of HITs completed
Information extraction (relation annotation)	56 ($0.10 per HIT)
RTE (language generation)	54 ($0.10 per HIT)
RTE (classification)	86 (25 x $0.25, 61 x $0.10 per HIT)
Paraphrase relations (classification)	270 (56 x $0.05, 56 x $0.10, 104 x $0.25 per HIT)
Total HITs completed	412

any higher than the agreement that is achieved at a pay per HIT of $0.05.

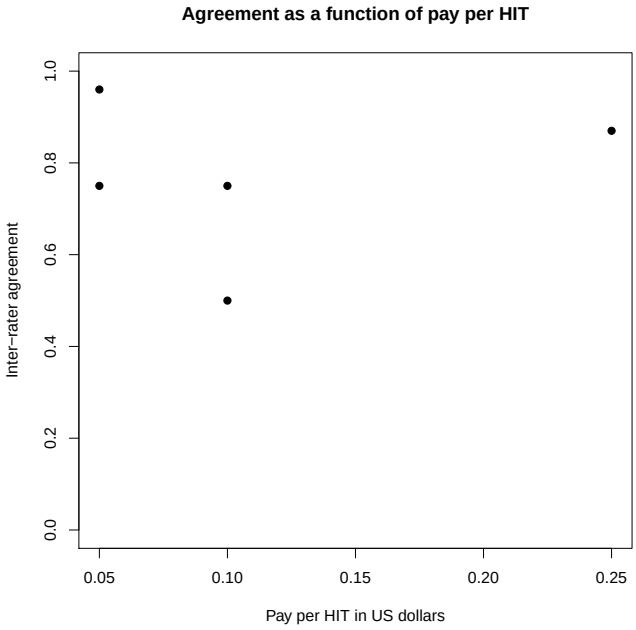

Figure 2: There is no relationship between the pay per HIT and the agreement achieved on a classification task. Each data point on the graph represents a separate set of tasks. 326 HITs were completed in the course of these 5 sets of tasks.

We cannot achieve higher agreement by paying more. Other than these two findings, the expectations listed in the Introduction were supported.

2.1. Why doesn't effective hourly rate increase as a function of pay per HIT?

Examining the time spent per HIT as a function of pay per HIT, we see why the effective hourly rate does not go up as the pay per HIT increases. Figure 3 shows the average time per task as a function of the pay per HIT. There *is* a linear relationship between the pay per HIT and the average time per assignment: as the pay per HIT goes up, the average time per assignment goes up. That is, the more the workers are paid, the more time they spend on each individual HIT. The correlation between them is very strong, at R = 0.92. Thus, even though the pay per HIT increases, the effective hourly rate stays about the same.

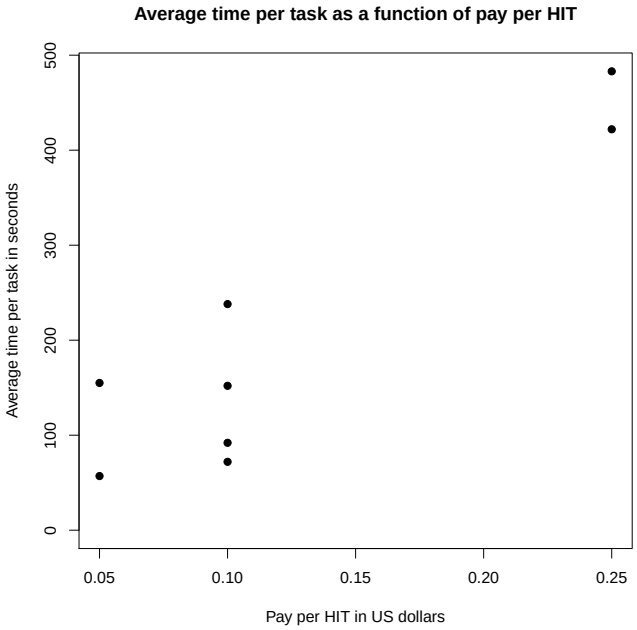

Figure 3: There is a linear relationship between the pay per HIT and the average time per task, R = 0.92. Each data point on the graph represents a separate set of tasks. 412 HITs were completed in the course of these 8 sets of tasks.

3. Discussion and Conclusions

3.1. Discussion

Although to the best of the authors' knowledge, the specific issue examined in this paper has not been studied before, there is a considerable amount of relevant work on the subject of crowdsourcing methods in general and crowdsourcing for linguistic resource creation in particular. (Callison-Burch and Dredze, 2010) give an overview of the results of a workshop on the use of Amazon Mechanical Turk to create data sets for natural language processing, held under the auspices of the Association for Computation Linguistics. The paper describes the results of 24 attempts to create language resources with Amazon Mechanical Turk, and gives some recommended practices for using the platform, including trying the task yourself and then having someone outside of the field try it, in order to assess the "doability" of the task and to estimate the time per HIT, in order to allow you to offer fair remuneration. (Sabou et al., 2012) point out some of the salutary effects of crowdsourcing linguistic resource construction, including diversification of

the task types, languages, resource types, and linguistic phenomena. In counterpoint, (Sagot et al., 2011) present a wide-ranging critique of for-pay crowdsourcing for language resource development in general, including observations consistent with the idea that crowdsourcing might not be as inexpensive as is widely assumed when one takes into account the costs of developing the interface, validating the data, and post-Turking processing; and the impossibility of determining with certainty the native language of Turkers. (Snow et al., 2008) measured the agreement between Turkers and expert annotators for five tasks, including recognizing textual entailment (the task type for 140 of the 412 HITs that were the source of the data in this paper). They found high agreement rates for all five task types. (Adda et al., Undated) also give a list of best practices, many of which deal with the ethical issues involved in crowdsourcing. These include taking into account the amount of time necessary to accomplish the task, including an estimated *hourly* wage in the work request (in addition to the pay per task that is automatically included by Amazon), defining in advance objective measures for deciding when work will be rejected (that is, not reimbursed) and making those measures known to potential Turkers, giving immediate feedback, and not requesting tasks anonymously.

The fact that ethical issues exist concerning the use of for-pay crowdsourcing comes up repeatedly in these papers. It is typical for those papers that recommend best practices for crowdsourcing recommend paying a fair rate. This does not seem like a controversial recommendation. However, the data presented here suggest that it might be more difficult to figure out how to do so than it appears at first glance—simply offering a higher pay rate per task does not result in a higher effective rate of pay.

3.2. Conclusions

We examined the relationship between the pay that is offered for each task on a crowdsourcing platform and the amount that a worker earns for performing that test. The data from eight sets of tasks comprising 412 HITs is consistent with the surprising finding that there is no relationship between them. Paying more per HIT does not cause workers to earn more per HIT: the higher the rate of pay, the more time workers spend on individual HITs. So, the effective hourly rate stays roughly the same: workers do not earn more regardless of how much we pay per HIT. This finding is consistent across a variety of NLP application data types (information extraction, recognizing textual entailment, and paraphrasing) and resource-building task types (classification and language generation). The finding has serious implications for language resource builders who want to behave ethically in their treatment of workers: other means besides higher pay per HIT must be found in order to compensate workers fairly. *The findings of this paper should not be taken as an endorsement of unfairly low pay rates for crowdsourcing workers. Rather, the intention is to point out that additional measures, such as pre-calculating and communicating to the workers an average hourly, rather than per-task, rate must be found in order to ensure an ethical rate of pay.*

4. Bibliographical References

Adda, G., Mariani, J.-J., Besacier, L., and Gelas, H. (2013). Economic and ethical background of crowdsourcing for speech. In *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*, pages 303–334. Wiley.

Adda, G., Mariani, J. J., and Besacier, L. (Undated). Analyse économique, juridique et éthique du crowdsourcing pour le TAL.

Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12. Association for Computational Linguistics.

Callison-Burch, C. (2014). Crowd-workers: Aggregating information across Turkers to help them find higher paying work. In *The Second AAAI Conference on Human Computation and Crowdsourcing (HCOMP-2014)*, November.

Chilton, L. B., Horton, J. J., Miller, R. C., and Azenkot, S. (2010). Task search in a human computation market. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 1–9.

Cohen, K. B., Pestian, J., and Fort, K. (2015). Annotateurs volontaires investis et éthique de l'annotation de lettres de suicidés. In *ETeRNAL (Ethique et Traitement Automatique des Langues)*.

Fort, K., Adda, G., and Cohen, K. B. (2011). Amazon Mechanical Turk: gold mine or coal mine? *Computational Linguistics (editorial)*, 37(2).

Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C., and Schader, M. (2011). Managing the crowd: Towards a taxonomy of crowdsourcing processes. In *AMCIS 2011 Proceedings*.

Ipeirotis, P. (2010). Analyzing the Amazon Mechanical Turk marketplace. CeDER Working Papers, http://hdl.handle.net/2451/29801, September. CeDER-10-04.

Ross, J., Zaldivar, A., Irani, L., and Tomlinson, B. (2009). Who are the Turkers? worker demographics in Amazon Mechanical Turk. Social Code Report 2009-01.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., and Tomlinson, B. (2010). Who are the crowdworkers?: shifting demographics in Mechanical Turk. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, CHI EA '10, New York, NY, USA. ACM.

Sabou, M., Bontcheva, K., and Scharl, A. (2012). Crowdsourcing research opportunities: lessons from natural language processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, page 17. ACM.

Sagot, B., Fort, K., Adda, G., Mariani, J., and Lang, B. (2011). Un turc mécanique pour les ressources linguistiques: critique de la myriadisation du travail parcellisé. In *TALN'2011-Traitement Automatique des Langues Naturelles*.

Silberman, M. S., Ross, J., Irani, L., and Tomlinson, B.

(2010). Sellers' problems in human computation markets. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 18–21.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.

Gratis, Libre, or Something Else? Regulations and Misassumptions Related to Working with Publicly Available Text Data

Jana Diesner, Chieh-Li Chin
The iSchool/ Graduate School of Library and information Science (GSLIS)
University of Illinois Urbana Champaign
501 E Daniel Street
Champaign, IL, 61820, USA
E-mail: jdiesner@illinois.edu, cchin6@illinois.edu

Abstract

Raw, marked up, and annotated language resources have enabled significant progress with science and applications. Continuing to innovate requires access to user generated and professionally produced, publicly available content, such as data from online production communities, social networking platforms, customer review sites, discussion forums, and expert blogs. However, researchers do not always have a comprehensive or correct understanding of what types of online data are permitted to be collected and used in what ways. This paper aims to clarify this point. The way in which a dataset is "open" is not defined by its accessibility, but by its copyright agreement, license, and possibly other regulations. In other words, the fact that a dataset is visible free of charge and without logging in to a service does not necessarily mean that the data can also be collected, analyzed, modified, or redistributed. The open software movement had introduced the distinction between free as in "free speech" (freedom from restriction, "libre") versus free as in "free beer" (freedom from cost, "gratis"). A possible risk or misassumption related to working with publicly available text data is to mistake gratis data for libre when some online content is really just free to look at. We summarize approaches to responsible and rule-compliant research with respect to "open data".

Keywords: open source data, user and professionally generated online content, gratis versus libre text data, ethics, data repositories

1. Introduction and Problem Statement

Raw, marked up, and annotated text corpora available to the research communities in Natural Language Processing (NLP), Computational Linguistics (CL), the digital humanities, and computational social science have enabled major progress and breakthroughs in these and other areas. Continuing to innovate requires access to contemporary text data that were generated by people using common information and communication technologies (ICT), such as data from online production communities (e.g., Wikipedia and GitHub), social networking platforms, customer review sites, discussion forums, and expert blogs. One problem with work in this area is that researchers do not always have a comprehensive or correct understanding of what types of user or professionally created web content are permitted to be collected and used in what ways (Kosinski, Matz, Gosling, Popov, & Stillwell, 2015; Vitak, Shilton, & Ashktorab, 2016; Zevenbergen et al., 2015; Zimmer, 2010). This paper aims to clarify this point. We focus on risks for researchers who gather and utilize content from publicly available sites rather than on privacy risks for people who make their information available online.

1.1 Benefits of Working with Publicly Available Text Data

On the beneficial side, working with data at any scale that were generated by people who use ICTs and who interact with others and with information within these infrastructures allows for considering both the content and structure of social interactions (Lazer et al., 2009) and for re-evaluating theories that are based on data generated in offline or non-ICT-facilitated environments (Diesner, 2015; Kleinberg, 2008). Research based on contemporary interaction and text has promoted the emergence and advancement of the fields of network science, web science and internet science (Tiropanis, Hall, Crowcroft, Contractor, & Tassiulas, 2015).

Recognizing these benefits, some members of the scholarly community and their funders have been advocating for open access to data, code, knowledge and publications (Hodgson et al., 2014). Corresponding legal and technical solutions have been developed. Examples include copyright licenses by the Creative Commons[1] and open source licenses for software (for an overview see Opensource.org), as well as repositories that enable reliable and persistent access to publications, e.g., PubMed[2] for biomedical literature, as well as to domain specific and general science data (for an overview see "Recommended Data Repositories," 2016).

1.2 Risks of Working with Publicly Available Text Data

On the controversial side, scholars and practitioners might have an unclear or incomplete understanding and different conceptualizations of what "open source data" means and what this meaning implies for their practical, day-to-day work (Diesner & Chin, 2016; Vitak et al., 2016; Zevenbergen et al., 2015). Reasons for this effect include changing norms and regulations over time, and insufficient training on this topic.

Ethicists and privacy scholars have long argued that

[1] https://creativecommons.org
[2] http://www.ncbi.nlm.nih.gov/pubmed

working with user created, publicly available data can involve privacy risks for the individuals who generated and publish these data (Daries et al., 2014; Hoffman & Bruening, 2015; Lane, Stodden, Bender, & Nissenbaum, 2014). Several check points and risk mitigation mechanisms have been put in place, such as updates to Institutional Review Board (IRBs) processes. However, data from online sources might not be subject to review by an IRB if the researchers did not interact with the subjects and the data were already publicly available. Furthermore, collecting and using data from online sources may conflict with other types of regulations, including copyright, terms of service, established cultures in research communities, and personal values (Diesner & Chin, 2015; Kosinski et al., 2015; Zevenbergen et al., 2015). Deviating from these norms and rules may entail risks for researchers, their institutions and scientific communities, and the reputation of science (Zimmer, 2010).

In the remainder of this paper, we first briefly review classic types of sources for text corpora and related regulations. We then clarify what "open source data" means in theoretical and practical terms, and discuss potential reasons for confusion. Finally, we outline possible approaches to the responsible and ethical conduct of research that involves publicly available text data.

2. Background: Sources and Related Regulations for Working with Text Corpora

Some of the resources that have been widely used in the NLP and CL communities were prepared for and released as part of competitions and associated professional meetings, such as the "Text Retrieval Conference" (TREC)[3], "Automated Content Extraction" (ACE)[4], and the "Message Understanding Conference" (MUC)[5]. These data and related evaluation metrics have been serving as acknowledged standards and benchmarks for developing and assessing new computational solutions. Much of this work has been initiated and supported by US-based, federal funding agencies, such as the National Institute of Standards and Technology (NIST). Some of these data are now administered, maintained and distributed by the Linguistics Data Consortium (LDC)[6].

Furthermore, long-standing academe-based initiatives and collaborations have resulted in curated repositories, codebooks, lexicons, and annotations for domain-specific text coding purposes, such as the Human Relations Area Files (HRAF)[7] for the field of cultural anthropology, or the former Kansas Event Data System (KEDS)[8] for political science (Gerner, Schrodt, Francisco, & Weddle, 1994; Schrodt, Yilmaz, Gerner, & Hermreck, 2008).

More recently, private-public partnerships have resulted in the release of large scale archives of digitized text data, such as the HathiTrust[9] (Christenson, 2011; Wilkin, 2009). Some of these data are annotated for various types of textual features, e.g., entities and relations in the "Global Database of Events, Language, and Tone" (GDELT)[10] (Leetaru & Schrodt, 2013).

Most of the mentioned as well as other data sources that are commonly used for NLP and CL purposes include copyright statements, license agreements, or terms of service statements that determine how the data can or must be obtained, managed and used. However, for the wide range of human generated, publicly available content in the form of unstructured (e.g., blog entries) and semi-structured (e.g., Wikipedia articles) text data as well as mixed data (e.g., text and images) that are not behind a pay wall or a login wall, researchers might have a less clearly defined understanding of ethical and rule-compliant practices for data acquisition and utilization.

3. Regulations for Working with Publicly Available Text Data

For the purpose of this paper, "publicly available" means that the data are not behind a pay wall or a login wall, and can be accessed by anybody with a web-enabled device. Furthermore, we divide "publicly available text data" (short PATD) into two groups. First, data provided by ordinary users who utilize ICTs to generate, post or publish information ("user generated web content"), which includes a wide range of social media data. Second, data generated by companies and professional or paid staff, such as online newspaper articles ("professionally produced web content").

In reality, things can be more complex: Some webpages provide both types of information, e.g., Amazon features product descriptions from commercial providers and user reviews of these products, and newspaper websites provide articles written by journalists which users can comment on. Other webpages display snippets of content that originates from other sites and providers; sometimes justifying this practice with the fair use portion of the copyright law.

The ways in which one can engage with either type of PATD are governed by multiple sets regulations, including (1) personal values and ethics, (2) norms and rules that may differ by institution, sector and country (e.g., IRBs or the "Health Insurance Portability and Accountability Act" (HIPAA), (3) copyright law (including fair use), (4) privacy regulations, (5) security regulations, (6) terms of service, and (7) technical solutions (for a brief overview see Diesner & Chin, 2016).

Understanding and implementing these rules can be complicated. Educating instructors and students on these topics may lag behind technical feasibility and reality. Some regulations keep emerging and are later adjusted;

[3] http://trec.nist.gov/

[4] http://www.itl.nist.gov/iad/mig/tests/ace/

[5] http://www-nlpir.nist.gov/related_projects/muc/index.html

[6] https://www.ldc.upenn.edu/

[7] http://hraf.yale.edu/

[8] http://www.aaas.org/page/kansas-event-data-system-keds-project

[9] https://www.hathitrust.org/

[10] http://www.gdeltproject.org/

making them moving targets. Some rules are explicit, while others are more tacit, such as personal values and expected culture in scientific communities. Also, some explicit rules, such as terms of service, might be difficult to translate into practical solutions. The resulting lack of clarity as well as instances of research that received controversial reactions (Kramer, Guillory, & Hancock, 2014; Zimmer, 2010) have stirred debates about responsible and ethical ways for collecting and using PATD (Vitak et al., 2016).

3.1 What are "Open Source Data"?

The way in which a dataset is "open" is not defined by its accessibility, but by its copyright agreement, license, and possibly other regulations. In other words, the fact that a dataset is visible free of charge and without logging in to a service does not necessarily mean that the data can also be collected, analyzed, modified, or redistributed (Zevenbergen et al., 2015; Zimmer, 2010).

The open software movement has introduced the distinction between free as in "free speech" (freedom to use, modify and redistribute information with little restriction, "libre") versus free as in "free beer" (i.e. freedom from cost, "gratis") (Lessig, 2004; Stallman, 2002). The risk with PATD is that gratis might be mistaken for libre when the data really just are gratis (to look at). This misassumption may due to a variety of reasons, such as insufficient expertise, evolving norms, or prior work (performed under different regulations) that has set an example.

That being said, some PATD truly are in the public domain (libre) because they have an open source license. For example, articles, talk pages, and structured meta-data from Wikipedia [11] are released under the Creative Commons Attribution-ShareAlike License[12], which allows people to copy, distribute, adapt and transmit the work as long as they attribute the work and publish any derivations under the same, similar or a compatible license. Another example is WordNet (Fellbaum, 1998), a widely used lexical database of terms and their relationships, which is provided under its own open source license[13]. Also, some text data provided by several US-based federal agencies are in the public domain as the content "was prepared by employees of the United States Government as part of their official duties and, therefore, is not subject to copyright"[14]. An example are transcripts of congressional hearings, which are available through the website of the General Publishing Office (GPO)[15].

However, a wide range of social media data (user generated web content), including posts on many product and film review sites, as well as regular media data (professionally produced web content), including the online presence of classic print media, are gratis for personal use but not libre. In either case, the terms of use for these data are typically defined by the owner of the website. Users who provide content on these sites agree to these terms as part of the process of releasing their work on them. In fact, much of the publicly available online content, especially (social) media data, are protected by terms of service. These terms are often presented as browse-wrap agreements at the bottom of a webpage. Via these agreements, content providers often grant webpage visitors the right to access and making personal, non-commercial use of the data. Overall, rules for interacting with online content can make their permitted use comparable to reading notes on a traditional bulletin board or looking through a store window (gratis).

4. Approaches to Responsible Research with Publicly Available Text Data

Rule-compliant research can be achieved in several ways. First, considering applicable agreements requires awareness and acknowledgement of their existence, and an understanding of their actionable meaning. This applies to both terms of service and other regulations that may apply, such as the "Fair Information Practice Principles" (FIPPs) [16] or the "Health Insurance Portability and Accountability Act" (HIPAA)[17]. Mastering this step is mainly a matter to education and experience.

Second, some data providers offer technical solutions that explicate or implement the sites' data access and sharing, e.g., mainly robot.txt files and APIs. Considering such technical solutions requires a certain level of proficiency.

Third, researchers can contact data providers to obtain permission for data gathering and use under certain conditions. This solution is limited in its scalability as it involves a certain amount of administrative overhead for both sides.

Fourth, while user generated content is still a fairly recent phenomenon and data source, and related policies and regulations are still being developed, some companies have emerged that act as brokers of data between (corporate) content providers and end users, e.g., Crimson Hexagon[18] and BrandWatch[19]. In exchange for a fee, such services typically offer their customers increased data access (fire hose) over public APIs (garden hose) as well as data analytics computed over the raw material. The revenue from these for-pay models is typically not directly shared with users who generated the content, but might be invested in sustaining and improving platforms, services, features, and user experience, for example.

Fifth, we suggest that a novel and alternative solution

[11] https://www.wikipedia.org/

[12] https://en.wikipedia.org/wiki/Wikipedia:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License

[13] https://wordnet.princeton.edu/wordnet/license/

[14] http://www.ntsb.gov/about/Policies/Pages/Policies.aspx

[15] https://www.gpo.gov/fdsys/browse/collection.action?collectionCode=CHRG

[16] http://www.nist.gov/nstic/NSTIC-FIPPs.pdf

[17] http://www.hhs.gov/hipaa/index.html

[18] http://www.crimsonhexagon.com/

[19] https://www.brandwatch.com/

would be to enable content generating users to opt in to having their data being freely (libre) used under certain conditions, e.g. demanding that de-identification is performed. This opt-in choice could be provided as part of the process of posting content online.

4.1 Consequences of Using Gratis but not Libre Text Data on Reproducibility

Finally, once a researcher has obtained user or professionally created text data from an online source, another issue with these data may arise. Research should be reproducible, which has already become increasingly challenging with dynamic data and tools (Stodden, Leisch, & Peng, 2014). Federal funders encourage the free (libre) sharing of data and code to enable the reproducibility of work and maximizing the benefits of investing tax payers' dollars. Multiple funding agencies have started to require data management plans as part of proposals submissions. In these plans, researchers are asked - among other criteria - to specify how they intend to provide the outcomes of their work after project completion. Analogously, university libraries, among other stakeholders, have started to create, curate and administer data repositories where researchers can upload and search for data. However, if the data are proprietary or protected in other ways, for example by copyright or terms of service, making them available might not be an option for researchers. For example, some social media data can be obtained in a permitted and lawful manner, such as tweets via the Twitter API[20] or information from certain Facebook pages through their API[21] (both services have increasingly reduced the data that ordinary people can obtain through the APIs, e.g., Twitter in terms of the time window into the past, and Facebook with respect to access to peoples' personal pages). Researchers have annotated such data for a variety of text characteristics, e.g., sentiment, opinions and factuality, often with the goal of building prediction models (McAuley & Leskovec, 2013; Pang & Lee, 2008). However, sharing (redistributing) the annotated (modified) data may not be permitted. Only providing pointers or unique key identifiers that link annotations to the original source can be one technical solution to this issue. Finally, prediction models built based on annotating such data may also be subject to inherited licenses and agreements, even though the original data cannot be reconstructed from these models.

5. Conclusion

In summary, the process of working with user and professionally generated, publicly available text data can be regulated by a multitude of rules and norms. Developing the awareness, knowledge and skills to responsibly consider these rules and account for grey zones is a challenging and evolving issue. One common risk is to mistake gratis data (access free of charge) as libre (collect and use with little or no restriction).

We believe that a vibrant dialogue between academe, the private sector and policy makers is needed to move ahead with establishing best practices and rules that enable the advancement of science, respect peoples' privacy, and offer incentives for commercial activities.

[20] https://dev.twitter.com/overview/documentation

[21] https://developers.facebook.com/

6. Acknowledgements

This work is supported in part by the FORD Foundation, grant 0155-0370, and a faculty fellowship from the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana Champaign. We also thank the reviewers for their comments.

7. Bibliographical References

Christenson, H. (2011). HathiTrust. *Library Resources & Technical Services, 55*(2), 93-102.

Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., . . . Chuang, I. (2014). Privacy, anonymity, and big data in the social sciences. *Communications of the ACM, 57*(9), 56-63.

Diesner, J. (2015). Small decisions with big impact on data analytics. *Big Data & Society, 2*(2).

Diesner, J., & Chin, C. L. (2016). *Seeing the forest for the trees: considering applicable types of regulations for the responsible collection and analysis of human centered data.* Paper presented at the Human-Centered Data Science (HCDS) Workshop at 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW), San Fancisco, CA.

Diesner, J., & Chin, J. C. (2015). *Usable Ethics: Practical considerations for responsibly conducting research with social trace data*. Paper presented at the Beyond IRBs: Ethical Review Processes for Big Data Research, Washington, DC.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Gerner, D. J., Schrodt, P. A., Francisco, R. A., & Weddle, J. L. (1994). Machine coding of event data using regional and international sources. *International Studies Quarterly, 38*(1), 91-119.

Hodgson, C., Suber, P., Kiley, R., Kaufman, R., Goodrich, J., Eve, M. P., . . . Sutton, C. (2014). Open access infrastructure: where we are and where we need to go. *Information Standards Quarterly, 26*(2), 1-14.

Hoffman, D., & Bruening, P. (2015). *Rethinking privacy: Fair information practice principles reinterpreted*. Paper presented at the 37th Annual International Data Protection and Privacy Commissioners' Conference.

Kleinberg, J. (2008). The convergence of social and technological networks. *Communications of the ACM, 51*(11), 66-72.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines.

American Psychologist, 70(6), 543-556.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, 111*(24), 8788-8790.

Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (2014). *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. New York, NY, USA: Cambridge University Press.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., . . . Gutmann, M. (2009). Life in the network: the coming age of computational social science. *Science, 323*(5915), 721-723.

Leetaru, K., & Schrodt, P. A. (2013). *GDELT: Global data on events, location, and tone, 1979–2012.* Paper presented at the ISA Annual Convention, San Francisco, California, USA.

Lessig, L. (2004). *Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity*: Penguin.

McAuley, J., & Leskovec, J. (2013). *Hidden factors and hidden topics: understanding rating dimensions with review text (RecSys)*. Paper presented at the Proceedings of the 7th ACM conference on Recommender Systems, New York, NY.

Opensource.org. Open Source Licenses by Category. Retrieved from https://opensource.org/licenses/category

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135. doi:10.1561/1500000001

Recommended Data Repositories. (2016). Retrieved from http://www.nature.com/sdata/data-policies/repositories#general

Schrodt, P. A., Yilmaz, O., Gerner, D. J., & Hermreck, D. (2008). *Coding sub-state actors using the CAMEO (Conflict and Mediation Event Observations) actor coding dramework*. Paper presented at the Annual Meeting of the International Studies Association, San Francisco, CA.

Stallman, R. (2002). *Free software, free society: Selected essays of Richard M. Stallman*: Lulu. com.

Stodden, V., Leisch, F., & Peng, R. D. (2014). *Implementing Reproducible Research*: CRC Press.

Tiropanis, T., Hall, W., Crowcroft, J., Contractor, N., & Tassiulas, L. (2015). Network science, web science, and internet science. *Communications of the ACM, 58*(8), 76-82.

Vitak, J., Shilton, K., & Ashktorab, Z. (2016). *Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community*. Paper presented at the 9th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2016) San Francisco, CA.

Wilkin, J. (2009). BackTalk: HathiTrust and the Google Deal. *Library Journal*.

Zevenbergen, B., Mittelstadt, B., Véliz, C., Detweiler, C., Cath, C., Savulescu, J., & Whittaker, M. (2015). Philosophy Meets Internet Engineering: Ethics in Networked Systems Research. (GTC Workshop Outcomes Paper): Oxford Internet Institute, University of Oxford

Zimmer, M. (2010). “But the data is already public”: on the ethics of research in Facebook. *Ethics and Information Technology, 12*(4), 313-325.

On the Need for a Global Declaration of Ethical Principles for Experimentation with Personal Data

Wessel Reijers[1], Eva Vanmassenhove[2], David Lewis[3], Joss Moorkens[4]

[1]ADAPT Centre/School of Computing, Dublin City University, Ireland,

[2]ADAPT Centre/School of Computing, Dublin City University, Ireland,

[3] ADAPT Centre at Trinity College Dublin, Ireland,

[4]ADAPT Centre/School of Computing, Dublin City University, Ireland

E-mail: wreijers@adaptcentre.ie, eva.vanmassenhove2@mail.dcu.ie, dave.lewis@adaptcentre.ie, joss.moorkens@dcu.ie

Abstract

In this paper, we argue that there is a growing need for a globally accepted set of ethical principles for experimentation that makes use of collections of personal multimodal data. Just as the Helsinki declaration was signed in 1964 to provide ethical principles that would guide all experimentation with human subjects, we argue that today the "digital personae" ought to be protected by a similar, globally endorsed declaration that informs legal regulations and policy making. The rationale for such a declaration lies in the increasing pervasiveness of the use of personal data in many aspects of our daily lives, as well as in the scattered nature of data research for which particular implementations of research ethics at the level of a single institution would not suffice. We argue that the asymmetry between ethical standards of public and of commercial entities, the borderless and boundless nature of online experimentation and the increasing ambiguity of the meaning of online "experiments" are compelling reasons to propose a global declaration of ethical principles for experimentation with personal data.

Keywords: Helsinki declaration, ethical principles, data collection

1. Introduction

After the grave atrocities of World War II with regards to the experimentation with human subjects, a first-ever internationally accepted code with ethical principles – the Nuremberg code – was drafted to protect human subjects in research settings on a global scale in 1947. In 1964, the Helsinki declaration – a successor of the Nuremberg code – came into effect and has since then been revised on multiple occasions (Carlson, Boyd, and Webb, 2004).

While the Helsinki declaration focuses on the treatment of the physical human being and is primarily concerned with experimentation in the medical sciences, contemporary life is increasingly shaped by the "digital persona" or "digital profiles" of people interacting with each other in cyberspace (Roosendaal, 2014). As Jouhki et al. (2015) argue, experiments with personal data challenge the idea of "personhood" in the context of research ethics. In this paper, we argue that this development, together with the diffuse understanding of "experiments" in cyberspace calls for a global declaration of ethical principles for experimentation with personal data that can inform legal regulations and policies.

2. The Rationale for a Global Declaration of Ethical Principles for Online Experimentation

Ross raises the question: "do research ethics need updating for the digital age?" (Ross, 2014). He argues that online experiments such as the controversial "Facebook experiment", as well as the emerging idea that anyone on the Internet is constantly subject to experiments, lead us to rethink what it means to experiment with human subjects. The Facebook experiment (Kramer, Guillory, and Hancock, 2014), which we will use in this article to illustrate core parts of our argument, was conducted by Facebook in collaboration with researchers of Cornell University. It manipulated the news feeds of 689,003 of its users and collected the corresponding user data that the Cornell University researchers used for further analysis. The purpose of the study was to investigate the effects of a more "positive" or "negative" newsfeed on the user's behaviours. The standard ethics review that would have applied to this study had it been entirely conducted by Cornell University did not apply because Facebook took care of the manipulation and collection of the personal data. This online experiment resulted in various academic discussions on the ethical implications of experimentation with personal data (e.g. see Flick, 2016; Kleinsman and Buckley, 2015; Jouhki et al., 2015).

Apparently, different standards for ethical conduct exist in our online and offline lives. Although it seems unimaginable that we would accept researchers influencing our emotional states by for instance offering us different kinds of psychological stimuli without explicitly asking our consent, this practice is both possible and seems to remain largely unquestioned with regards to our online interactions. One could argue that simply the same ethical standards should simply apply to both online and offline studies. Yet, however, it seems incorrect to equate harm that can be done to the physical subject with harm that can be done to the digital subject. Therefore, even though experiments with personal data

can lead to kinds of harm, it seems that existing policies and ethical frameworks cannot simply be assumed to directly be applicable to them (Flick, 2016). We argue that new experimental practices that involve the use of collections of personal, multimodal data lead to three core dilemmas for research ethics of data science:

1. The blurring character of the "data researcher"
2. The borderless and boundless nature of digital experiments
3. The ambiguity of "harm" caused by online experimentation

We argue that a partial solution to overcome these dilemmas would be a global declaration of ethical principles that guides practices of experimentation with personal data. In the following sections, we support our argument by focusing on the three dilemmas.

2.1 Who is the Data Researcher?

First, there seems to be an asymmetry between those researchers referred to as "data scientists" (Metcalf, 2015), who usually work at public research institutions and data researchers working in commercial environments. In the case of the Facebook experiment, the usual constraints that would apply to the work of data scientists, such as the requirement to ask the participants for explicit informed consent to manipulate their user experience (the kind of newsfeed they would see), was circumvented due to the involvement of a commercial entity that was said to have independently performed the data collection (Flick, 2016).

A considerable number of initiatives have been developed to harmonise ethical principles and guidelines for data research (e.g., see: European Commission, 2013), but those are principally intended for implementation by public research institutions and do not - or only partially - apply to research conducted by commercial entities. Notable limitations to experimentation practices that such harmonised guidelines articulate are requirements for explicit, often written consent of research participants and the obligation for the researcher to consider the possible effects of their research on vulnerable groups. However, data research is often conducted by both public and commercial entities, as we can for example show to be the case for research on translation memories and machine translation, since it both includes the research communities of private companies and of researchers in university contexts (Pym, 2011). Thus, knowledge gained from online experiments in such data research communities originates not just from public research institutions, but also from commercial entities.

With regards to the use of personal data at public institutions - for instance the use of video material for emotion detection by machine learning algorithms – increasingly strict ethical procedures apply. The human subjects being filmed usually have to be informed about the methods and aims of the research and need to offer explicit consent for the fact that they are part of a scientific research. For commercial entities, for instance a data company analysing the response patterns of users showing different videos from an online video service, such ethical concerns typically do not apply. As Jouhki et al. (2015) argue, ethical concerns such as informed consent are in such cases often part of the Data Use Policy (Jouhki et al., 2015).

According to Flick (2016), while for public research institutions even the use of simple experimental tools such as questionnaires is "highly regulated"; commercial or collaborative research that makes use of experiments with personal data is barely regulated. Moreover, mere disclosure of the relevant information through Data Use Policies is usually considered as sufficient for achieving informed consent, despite arguably being ethically insufficient. We argue that this asymmetry between public and commercial experiments with personal data is a problematic one, for it instantiates to some extent an "ethical" and an "ethics-free" zone in research communities without proper justification. Moreover, when ethical concerns are taken into account by some researchers but not by others, it seems superfluous to implement forms of research ethics for data researchers. Just as medical researchers in private clinics have to comply with the same basic ethical principles as their colleagues in publicly funded institutions, so should it be the case for all data researchers that use personal data for experimentation purposes. A global declaration of ethical principles for data research could at least partially solve this problematic asymmetry.

2.2 Borderless and Boundless Experimentation

Second, we argue that the - to a certain extent - borderless and boundless nature of online experimentation practices calls for a global declaration of ethical principles for data research. Geographic boundaries of national jurisdictions are challenged by the absence of geographic borders in cyberspace (Drezner, 2004). Because digital content flows through a borderless cyberspace, meaning that online spaces could be inhabited by people from many different countries and cultures with no restrictions on their interactions, national or regional regimes of ethical principles for data research seem inadequate for dealing with the ethical concerns of experimentation with personal data. For instance, the Facebook experiment could have included many different users from different nationalities. In that case, it would not have been sufficient for an EU citizen to be protected by ethical guidelines for data research in an EU context, for (s)he could still be included in online experimentation practices conducted by entities that have their legal basis outside of the EU jurisdiction. For this reason, ethical guidelines for experiments involving personal data should be agreed upon and implemented at the global level, meaning that as many countries as possible should endorse a joint declaration on this matter.

Moreover, we have to deal with the "boundless" nature of experimentation with personal data, by which we mean the increasing potential for multiple ways in which such experimentation could take place (considering advances in the creation of tools for data-analysis). Even though the Facebook experiment set an important precedent for the understanding of experimentation with personal data, it does not represent the *only* type of experimentation that could take place by making use of personal data. For instance, experiments might make increasing use of multimodal data, by analysing voice recordings and video recordings to capture bodily

expressions and affective voice patterns. Such types of experiments might deal with even more intimate phenomena and might therefore give rise to more serious ethical concerns. Different manipulated multimodal stimuli might be presented to users: not just manipulated timelines as in the case of Facebook but for instance manipulated video feeds or manipulated communications between users. This development could lead to a so-called "Collingridge dilemma", which means that future impacts of new technologies or techniques cannot be easily predicted and that it is difficult to control or change the causes of impacts once a technology is entrenched (Tannert, Elvers, and Jandrig, 2007).

2.3 Potential "Harm" by Online Experimentation

Third, we need to discuss whether we can even speak about a potential harm for "participants" in online experiments that is similar to the harm that a declaration of ethical principles such as the Helsinki declaration is aimed at preventing. Full-scale online experiments based on collection of personal data are a relatively recent phenomenon and there seem to be good reasons to argue that "harm" in such experiments is significantly different from harm in for instance medical experiments. The way that the Facebook experiment manipulated the emotions of its users could be justifiable from a consequentialist point of view; which could be that the little harm done to some of the participants resulted in a greater user experience, and thus benefit, for all Facebook users (see Meyer, 2014). However, we need to take into account that the actual harm suffered by some of the participants is very opaque; we would have no means to assess what harm would be caused by this experiment for we have no data about the psychological or perhaps physical effects of the experiment on its participants. Moreover, the contents of some experiments seems to enable more grave concerns, for instance when concerned with the manipulation of political views of participants (Angelica and Fong, 2012).

One of the arguments raised in defence of Facebook's experiment stresses that the impact of the manipulation of the timelines of users was relatively limited and that Facebook alters it algorithm "all the time" (Meyer, 2014). However, these are exactly the issues that can raise concerns. How much could the risk for harm by experiments with personal data increase because of the increasing potential of instruments for data manipulation and analysis? And how can we distinguish between a distinct experiment with personal data and the day-to-day business of a "service that carries unknown emotional risks" (Meyer, 2014)? Because all processes of personalisation – the tailoring of digital contents to the needs of a human user (Yalcinalp and Gulbahar, 2010) - are in a certain way "experimental", in the sense that they measure certain variables of the behaviour of human users and use the outcomes of these measurements to construct generalisations, experimentation is an integral aspect of online interactions.

Such concerns call for a clear delineation of what it means to conduct an experiment with personal data. We might do this by setting certain criteria for calling interference in online interaction an experiment. For instance, whenever the conditions of online interaction are manipulated for a subset of people using a service, one might call it an experiment. Also, we might take into account the *kind* of manipulation in order to define experimentation with personal data. For instance, when a manipulation is strictly functional (e.g. changing a user-interface), one might refrain from qualifying it as an experiment, but when a manipulation is aimed at offering different experiences of affective contents, one might consider it as an experiment. In any case, we might want to assign a different meaning to "experiments" conducted in cyberspace than to those conducted in conventional research settings. A clear definition of an online experiment, and of the corresponding criteria to determine what is and what is not an experiment with personal data, should form the basis of the process leading to a global declaration for ethical principles that would apply to such experiments.

3. Conclusion & Discussion

In this paper, we argued for the need of a globally acknowledged declaration of ethical principles for experimentation based on personal data collections. We did so by discussing three dilemmas that are caused by the emergence of online experimentation practices with personal data. First, we argued that since two different ethical regimes exist for similar research – in the private and the public realms – we are in need of unified basic ethical principles that all researchers should comply with. Secondly, we argued that since Internet research is conducted in a borderless cyberspace, we are in need of a set of globally accepted ethical principles. Thirdly, we discussed the ambiguity of online 'experimentation' and ways in which we could reach a better understanding of this notion when considering the harm it could cause.

How could a global declaration of ethical principles eventually influence the practical reality of experimentation with personal data? First of all, it could inform guidelines for ethical conduct that apply to experimentation with personal data conducted by both public and commercial entities. Secondly, it could influence policies aimed at licencing organisations for conducting experiments with personal data. Licences could apply on the condition that certain ethical principles are respected. Similarly to licencing for drug companies, violation of ethical principles could lead to the cancellation of a licence for an organisation and consequently to the discontinuation of experimentation practices by that organisation.

Even though a global declaration of ethical principles for experimentation with personal data would be a step in the right direction, it would not be sufficient for resolving all of the concerns we have raised. In line with Brey (2012), we argue that it is crucial to also anticipate ethical impacts of emerging information technologies. Especially since biometrical data is increasingly integrated with our online existence, the firm line between our biomedical, physical selves and our online, digital selves seems to be blurring and therefore we need to pay equal attention to physical harm in the offline world as to "digital harm" in the online world. Therefore, next to setting ethical standards for current online experimentation practices, we should also look at ways in which we can anticipate and assess ethical impacts of emerging technologies.

4. Acknowledgements

This work has been supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centers Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund, by the European Commission as part of the FALCON project (contract number 610879), and by the Dublin City University Faculty of Engineering & Computing under the Daniel O'Hare Research Scholarship scheme.

5. References

Angelica, M. D, and Fong, Y. (2012). "A 61-Million-Person Experiment in Social Influence and Political Mobilization." *Nature* 489 (7415). doi:10.1016/j.surg.2006.10.010.Use.

Brey, P. (2012). "Anticipating Ethical Issues in Emerging IT." *Ethics and Information Technology* 14: 305–17. doi:10.1007/s10676-012-9293-y.

Carlson, R. V., Boyd, K.M. and Webb, D.J. (2004). "The Revision of the Declaration of Helsinki: Past, Present and Future." *British Journal of Clinical Pharmacology* 57 (6): 695–713. doi:10.1111/j.1365-2125.2004.02103.x.

Drezner, D. W. (2004). "The Global Governance of the Internet: Bringing the State Back In." *Political Science Quarterly* 119 (3): 477. doi:10.2307/20202392.

European Commission. (2013). *Ethics for Researchers*. Luxembourg: Publications Office of the European Union.

Flick, C. (2016). "Informed Consent and the Facebook Emotional Manipulation Study." *Research Ethics* 12 (1): 14–28. doi:10.1177/1747016115599568.

Jouhki, J., Lauk, E., Penttinen, M., Rohila, J., Sormanen, N., and Uskali, T. (2015). "Social Media Personhood as a Challange to Research Ethics." In *Successes and Failures in Studying Social Media: Issues of Methods and Ethics*. University of Jyvaskyla.

Kleinsman, J., and Buckley, S. (2015). "Facebook Study: A Little Bit Unethical But Worth It?" *Journal of Bioethical Inquiry*, 179–82. doi:10.1007/s11673-015-9621-0.

Kramer, A., Guillory, J.E. and Hancock, J.T. (2014). "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks." *Proceedings of the National Academy of Sciences* 111 (24): 8788–90. doi:10.1073/pnas.1412469111.

Metcalf, J. (2015). "Human-Subjects Protections and Big Data : Open Questions and Changing Landscapes," no. 2012: 1–13. http://bdes.datasociety.net/wp-content/uploads/2015/07/Human-Subjects-Lit-Review.pdf.

Meyer, M. N. (2014). "Misjudgements Will Drive Social Trials Underground." *Nature* 511 (7509): 265–265. doi:10.1038/511265a.

Pym, A. (2011). "Democratizing Translation Technologies–the Role of Humanistic Research." *Luspio Translation Automation Conference, Rome*, no. April: 12.

Roosendaal, A. (2014). "Digital Personae and Profiles as Representations of Individuals." In *Privacy and Identity Management for Life*, edited by Bezzi, M., Duquenoy, P., Fischer-Hubner, S., Hansen, M., and Zhang, G. 226–36. Springer.

Ross, M. W. (2014). "Do Research Ethics Need Updating for the Digital Age?" *American Psychological Association*. http://www.apa.org/monitor/2014/10/research-ethics.aspx.

Tannert, C., Elvers, H.D. and Jandrig, B. (2007). "The Ethics of Uncertainty." *EMBO Reports* 8 (10): 892–96. doi:10.1038/sj.embor.7401072.

Yalcinalp, S., and Gulbahar, Y. (2010). "Ontology and Taxonomy Design and Development for Personalised Web-Based Learning Systems." *British Journal of Educational Technology* 41 (6): 883–96. doi:10.1111/j.1467-8535.2009.01049.x.

Diffusion of Memory Footprints for an Ethical Human-Robot Interaction System

Agnes Delaborde[1] **and Laurence Devillers**[1,2]
[1]LIMSI-CNRS, Université Paris-Saclay, Orsay, France
[2]Université Paris-Sorbonne IV, Paris, France
{agnes.delaborde, laurence.devillers@limsi.fr

Abstract

Along with the constant amelioration of the artificial intelligence skills in robots, has arisen a strong will among the community to define ethical limits to the behaviors of the robots. The implementation of ethics and morality in an autonomous system represents a research challenge, and several practical propositions have been offered by the community. The authors propose trails for a Human-Robot Interaction (HRI) architecture in which the selective diffusion of footprints and logs extracted from the robot's memory (low-level inputs, interpretation, decisions, actions) would improve the traceability of the robot's internal decision-making, which could for example offer a guarantee of transparency in case of faulty or contentious situations. The description of the proposed architecture is based on the authors' studies on social Human-Robot Interaction systems designed in the context of the French robotic project ROMEO. The authors' proposition will be subsequently assessed in the course of a French transdisciplinary project involving the fields of robotics, law and artificial intelligence.

Keywords: Human-Robot Interaction, Robot ethics, Roboethics, Memory, Traceability

1. Introduction

The designing of a Human–Robot Interaction (HRI) system for an assistive robot requires carrying out data collections and experiments in social and assistive tasks, which will allow assessing that the robot proves to be technologically efficient, socially acceptable, and that it has a positive impact on the user's well-being. These three fundamental dimensions, notably summed up by (Feil-Seifer et al., 2007), are generally considered as a basis for the evaluation of the overall efficiency of an assistive robotic system.

An increasing interest for ethics has developed in the HRI research communities, leading to many works and reflections around what a robot should be allowed, or not allowed, to do when interacting with a human. Far from an exhaustive list, one can cite several works on ethics of robotics (Lin et al., 2011; Asaro, 2006; Capurro et al., 2009; Tamburrini, 2009). All these studies agree on the need to design the system accordingly with axiological considerations, thus endowing the system either with rules created with a moral conscience from the designers, or even a capability for moral discernment in the robot. Roboethics require a collaborative work with all the disciplines involved directly in the design of the robot (computer sciences, sociology, medicine...), but also from scholars likely to assess the integration of robots in society (philosophers, lawyers, economists). The importance of the transdisciplinarity in roboethics is for instance addressed in the workshop "The emerging policy and ethics of Human-Robot Interaction" @HRI2015 (Riek et al., 2015).

Although some of these considerations have been considered anticipatory for some years, they are today close to technological capacities of HRI systems, and should be at the heart of the designing process. Thus, an ethical awareness could translate into endowing the robot with the ability to make some of its internal data explicitly available, either to the user, to a nursing auxiliary in a nurse–user–robot triad, or to a legal expert in the case of a search for liability. The collection of personal data gives rise to many ethical and legal issues, one of these being that the individual has to be aware of what has been collected upon him(her)self; he(she) also needs to know the recipient and the purpose of the data collected.

This reflection takes place in the framework of the French projects Romeo and Romeo2[1] aimed at the design of an assistive robot, and in the French project TE2R "Traces, explication et responsabilité du robot" ("Footprints, explanation and liability of the robot")[2], a robotics and law collaboration meant to analyze in which way the behaviors of the robot can be tracked and explained, so as to facilitate a search for liability.

The authors will present in a first part of this study an overview of HRI research works that consider ethical issues in their models and implementations, and a overview of the way data can be processed, stored and used in a robotic memory. In a second section, the authors shall present the architecture of the HRI system designed in the framework of the project Romeo, which processes audio signal data so as to infer the user's emotional state and profile; they will highlight the different levels of memory footprints and decisions made by the robot that could be broadcast. The final section will concern the selective diffusion of the memory information according to the recipient (user, medical and legal experts) and several study cases. The authors will conclude by pointing out the necessity to find ways to inform the user about the way the robot processes his(her) personal data, so as to increase the user's trust in the robot.

[1] BpiFrance and Cap Digital, www.projetromeo.com
[2] Lidex Paris-Saclay and Institut Société Numérique

2. State of the Art

2.1. Ethics in Human–Robot Interaction Systems

Although works namely dedicated to recommendations for roboethics are legion, practical solutions to endow robots with a sense of basic axiology (allowing the robot do distinguish autonomously if its decisions are "fair", "moral") are still scarce in the community. Nonetheless, some highly interesting models have emerged. Notably, one could cite Wilson and Scheutz' recent proposition to tag each robotic action with a moral expectation score (Wilson and Scheutz, 2015), or van Wynsberghe's works which offers a methodology for a step-by-step ethical design of healthcare robotic systems (van Wynsberghe, 2013). The notion of morality is nevertheless a highly subjective and cultural issue, and implementing a really non-equivocally moral robot seems to be a extremely ambitious (yet highly interesting) scientific research topic.

Literature in robotics offers a really wide panel of advanced and practical reflections in the domain of emotional, social and assistive robotics, meant to increase the acceptability of the systems, and to respond efficiently to the needs of dependent populations. The authors shall cite for example (Sharkey and Sharkey, 2012) in their analysis of ethical issues in the domain of robotic assistance for the elderly, or also (Malle, 2015) whose study focuses on the moral competence of robots. Many works deal with the acceptation of assistive robots: adaption of its speed to its follower (Fiore et al., 2015), or the preservation of social distance (Correa et al., 2014), the acceptability of dog-like behaviours (Lakatos et al., 2013). The enhancement of the robot's acceptability and social efficiency is a particularly active field in robotics, which fosters many practical solutions. Although most of them are not presented namely as ethical, the reflection underlying these works is fully in the scope of roboethics, which consists in designing systems that can *help* users, and meet their specific needs (be them in terms of respect of the individual, or preservation of their ability to remain at home instead of going to special institutions, etc.).

In the context of the European Robolaw project, (Bertolini and Palmerini, 2014) broach the possibility of resorting to "black boxes" to identify the robot's inputs and resulting decisions, in a process of keeping the user informed about the data processing. In this present study, the authors will address the possibility of endowing a HRI interface with such a black box system, which could allow to broadcast (in real time or through logs) the salient information processed by the robot.

2.2. Robotic memory

In HRI, endowing the robot with a perception of its environment relies on the capture, and on the interpretation of the data produced by this environment. Data processed in the system can be presented on a continuum of levels of abstraction. On one extremity, one can find raw data (unmodified by the considered system), and on the other extremity data resulting from one or several processes of interpretation.

In itself, raw data is of little interest for a system supposed to react to its environment. It is necessary to encode it, to make it understandable and processable by the system. Designing a learning model consists in selecting specific features of the input data, potentially transform them (merging, classifying), so as to produce a new output data. For example, an audio signal can lead to the production of an emotional label.

Many research teams in HRI look into the most appropriate way to process the system's external perceptions, along different approaches: the enhancement of technologies for the capture of external data (microphones, cameras, noise filters...), automatic detection of phenomena (presence of the user, localization, speech, emotion...), and resulting interpretation on a higher level (lexical content, emotional profile, identity, intentions...).

Currently, the modeling of the robotic memory is mostly inspired by living beings' systems. Robots capture information about the environment (video, audio, haptic captures, system information), interpret them in terms of representations (user identity, timestamps, localization of the robot in a room...) and store them in memory through human mnesic mechanisms. In particular, designers draw their inspiration from the human explicit memory, which processes and stores the data relating to facts and events (episodic memory), and world knowledge and the semantic relations between the memorized elements (semantic memory). In (Pointeau et al., 2013), the robot acquires experience and stores it in an autobiographic memory (relational databases representing the episodic and semantic memories), which can be used notably to simulate the results of the robot's actions. (Stachowicz and Kruijff, 2012) model an episodic memory which allows the memorization and recollection in a robotic cognitive system. (Kasap and Magnenat-Thalmann, 2010) offers a model of episodic memory integrated in a decision-making module for a long-term affective interaction.

In this study, the authors do not focus on memory in terms of mnesic mechanisms, but along the way raw data can be interpreted and made available for the robot's behavorial decision-making. In Pointeau et al.'s work cited before, for example, the systems processes what the authors call a snapchat of the environment (date and time, type of action performed, semantic role of the object on which the action is performed – a part of this data is given by the user). This data can be considered as raw data which the mnesic system receives as inputs and stores in the episodic memory without any modifications. The semantic memory is build up from data classified along several levels, among which spatial, temporal, contextual. Automatic reasoning mechanisms extract regularities from the episodic memory: this memory is thus composed of interpreted data, which is the basis for the action selection of the robot.

3. System features

3.1. Context

The authors' research focuses on the automatic selection of the action of the robots, based notably on an automatic user profile derived from speech emotional cues (Delaborde and Devillers, 2010). The profile allows the robot to react according to a conceptual and overall representation of the user (an emotional profile), rather than on an ad hoc basis (an emotion).

3.2. Interaction scenarios

In previous studies, the authors examined several parameters of the interaction loop through experiments featuring potential end-users interacting with the robot Nao in the framework of the project Romeo. The authors notably carried out data collections in the context of daily assistance for adults suffering from a loss of autonomy, and with children playing a game with the robot (Delaborde and Devillers, 2012).

In the tested scenarios, the robot presented desirable and non desirable social behaviors which were tagged along the interpersonal circumplex. The circumplex, defined in several early sociological works (Strong et al., 1988; Leary, 1958) and subsequently massively used in artificial intelligence, presents the complete range of interpersonal positions of an individual, on two principal axes: above/below the interlocutor and opposed/together. The authors obtained several results about the emotional reactions of these two populations of users in the course of interactions with the robot Nao .

These two experiments allowed the authors to select, for each population and their respective interaction context, the set of behaviors which had the most positive impact on the users. Indeed, the authors observed, in the context of assistance for impaired users, that undesirable social behaviors triggered emotions that were less positive than what was expected by the scenario lines. They also noticed that in a context of children at play, behaviors which were tagged as undesirable did not really affect the positivity of the children, but were on the contrary perceived as funny and engaging.

3.3. System data

The system processes nonverbal paralinguistic inputs expressed orally by the user, so as to detect his identity and his affective expressions. As described in (Devillers et al., 2015; Tahon and Devillers, 2016), low level cues can be computed from the speech signal: duration of speaker turns, F0, energy, and other acoustic coefficients. Markers can be derived from machine learning techniques such as support vector machines trained on various statistical functionals and transformations applied to specific features of the signal. This allows endowing the system with emotional information such as an emotion label, an activation level (the strength of the emotion), the emotional valence (positive, negative, neutral) and laughter. It also conveys the duration of speech, and the duration before the speaker starts to talk to the robot. In summary, the emotional inputs used are the emotion label (anger, joy, happiness, neutral), the activation (high, low), the valence (positive, negative, neutral),the laughter (presence, absence), the time elapsed before the speaker's start of speech, and the duration of speech.

In the manner of the human memory system, the robotic system deals with a working memory that organizes the input data in a processable way. It produces a vector for each speaking turn composed of the identity of the speaker, the class of the sound (laughter, speech, robot's own voice), the duration of the turn, and emotional data of speech turns. The information is then processed by the memory and added to a pile, but also merged in terms of means and modes to get an overall representation of the emotional behaviors of the user. The activity of the robot (its behaviors) is also stored. These elements of information are the basis for a decision making, either in terms of action taking (expressing a behavior) or for an update of the user profile. The decision-making relies on a fuzzy system, where inference rules are constituted from psycho-sociological studies, as well as qualitative and quantitative studies carried out on corpora collected by the authors (Delaborde and Devillers, 2012). The diagram in Figure 1 presents the architecture of the data processing chain: audio signal is captured; a detection module provides paralinguistic data which constitutes a vector; data is stored in memory either as a history or merged data; decision-making processes the memory content so as to select a robot behavior or update the user's profile.

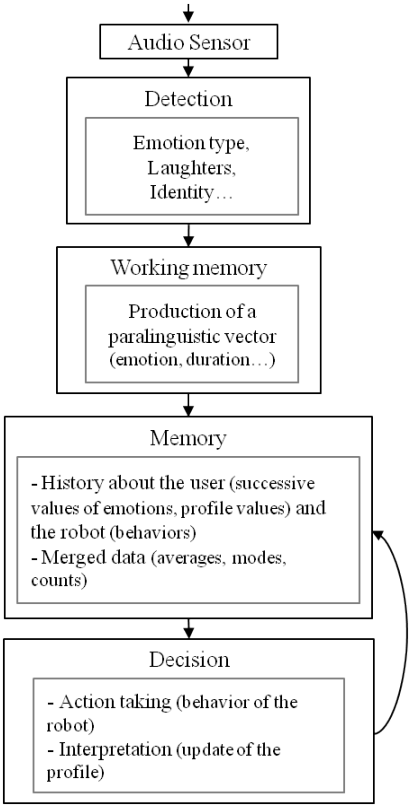

Figure 1: Processing of data and decision-making in the robotic system.

4. Selective diffusion of information

When a system collects data about the habits of the user – for example browser history, mails, etc. – so as to offer "customized content", this gives rise to obvious ethical issues. Indeed, even if the user has granted his(her) consent – he(she) would not be able to use the service otherwise – there is no way to determine exactly what personal data is being collected, how it is processed, and what interpretation it leads to: the user has no control. In these conditions, trust cannot set in.

In Human-Robot Interaction, detecting and understanding the emotions expressed by the user is crucial, insofar as this allows to react more naturally to the user. However, it is also compulsory that the user trusts his(her) robot, so as to accept its presence in his(her) domestic environment. The data processed and stored in the emotional system described in the last section constitutes footprints that the robot could broadcast to the user. This transparency could increase the level of trust and acceptation: the robot lives in my house, potentially records many personal details about me, but I know what it records and what it does with this data.

The selection of the broadcast information could be performed according to the recipient's nature. The authors will distinguish between the user, who would require data linked to the interaction and to the current task that the robot performs for/with him; the medical expert, for example in a triadic context of assistance, where the useful data would concern the assistance task and the well-being of the user; and also the legal expert who would search for the liability in the case of damage.

4.1. The user in a daily context

Broadcasting the information processed by the robot could increase the acceptability of the robot's behaviors, by offering a feedback about what it understood from the user, and the nature of its actions. However, the information made available could be limited to memory and decisional data, such as the interactional elements (the user's profile, the robot's chosen social attitude) or the data linked to the current task (in the case of schedule management for example), so as not to overwhelm the user with useless detailed information.

For example, the robot could inform that it has detected a state of "Anger" in the user, and that, since it is not usual in the user's temper (emotional profile), it is going to offer the user to play a game (or discuss, engage in a pleasurable activity...). In this way, the robot is clear about its reasoning, and, without being provided with all the details of the reasoning, the user can fill in the decision-making part that potentially has been performed. Research work has to be carried out so as to assess the acceptability and the added-value of such a system, and the effect upon the user if the robot produces conclusions that the user does not judge as rational.

The formulation of the elements displayed should be carefully selected. A robot cannot be allowed to tell a user that he(she) is not optimist at all. This piece of information can be crucial for a social robot which is sensible to its user's state, and scientifically justifiable, but can by no means be expressed as such. The process of popularization of science is directly concerned in this context: concepts as they are handled in science, without any further clarification, can lead to doubts and fears in the non-specialists.

4.2. The expert in a triadic context

In an assistive context, for example with a medical expert in a health-care triad, the robot could transmit data linked to the task and the role of the robot, so as to help in the monitoring of the patient. The robot, if it is set in the patient's room for example, would need also to make this information available to the user, to allow him(her) to know the content of the data it transmits to a third party. This would take part in establishing a positive relationship between the robot and the user. These information could be social information such as the one described in the previous section, but also the health record that the robot could make available to the nurse (frequency of pills taking, doctor appointments, etc.).

For example, the robot reminds the user that it is time for him to take his pills. The user refuses and expresses anger. The robot knows that this user is not used to losing its temper, and usually takes in pills on time. The robot could raise a warning and resort to the auxiliary nurse's help. The latter could decide to track the event and data that would have led the robot to call him. He could thus have access to data, on a black box principle, allowing him to know exactly what happened (the vectors in the working memory), and the decisions made consequently by the robot. The diffusion type would be similar to the case described in the previous sections, with the difference that the stored data would only be linked to situations tagged as relating to the assistive task: in this example, one can see that some personal information about the user (his/her temper) needs to be broadcast for the nurse to understand the situation. This highlights the vast difficulty of producing general rules about the limitations of the type of data that can be transmitted to third parties in the user's environment. While it can be easy to draw the outlines of rules from ad hoc and straightforward example (for example: the robot cannot transmit the information that the user made a phone call to his/her secret lover), endowing the robot with the capacity of automatically tagging what is really personal, private, and sensitive data may be subject to the same difficulty as tagging unequivocally the morality of actions.

4.3. The legal expert for liability establishing

In the case of situations causing damage, it could be conceivable to provide richer logs to allow and track the reasons for the damage, and eventually make a search for liability. These logs could contain chronological data about the working memory, the values in memory (raw and/or merged) and the decisions made. This diffusion would preserve the secrecy about the algorithms implemented in the

robot, while still providing sufficient information.

This aspect is being looked into in the framework of the project TE2R, and will be the subject of an upcoming collaborative publication. One ethical issue addressed in the context of a search of liability lies in the fact that the logs should be secured, to preserve the probative value of the data. This could imply automatically transferring the data to a secured external server, which would naturally raise several issues about the privacy of the data: who would store this data? In what country is stored my personal data, under which data protection rules? Who would have a scrutiny right upon this data? Legal and ethical aspects in robotics gives rise to numerous questions.

5. Conclusion and Discussion

The ethics in the processing of data is an intricate subject, torn between technical imperatives and considerations on the respect of the individual. The authors offers a way to improve the traceability of the robot's memory footprints and decisions, by designing a system that can broadcast its internal information according to the recipient. The data could increase the acceptability of the robot in the user, by giving a feedback about its own internal reasoning, or provide task-related information to assist the health-care nurse in the monitoring of a patient. The structured information could also be the basis for a legal search for liability. This last proposition will be assessed in the context of the TE2R project, to establish the judicial value of the logs, and the nature of the data that they should display.

It is noteworthy to mention that the feasibility of the diffusion of memory footprints relies on the physical design of the robot. Indeed, informing in real-time the user about the robot's internal changes would hinder a natural social communication, by flooding the user with information. This would be particularly true if the robot is only endowed with gestural and verbal means of communication: gestures allows only a low level of precision about the nature of the data transmitted, and the verbal channel would already be occupied with social and task-related subjects. One interesting way of making the data available to the user could be through the use of tablets (like the robots Synergy Swan or Pepper which integrates tablets in their design), either passively by continuously displaying data or, why not, through an active decision of the user to look into the robot's memory at specific times.

This present study addresses several major ethical issues, such as the question of the privacy of the data, and its personal nature: To what extent should a robot keep a log of the events that occurs at its owner's home? To what extent can the robot infer about the user? There are no definitive and easy to implement solutions to these issues. The integration of robots in the society is subject to many issues and fears among the public of non-specialists, sometimes aroused by mass media and science-fiction. The scientific community's ethical objective mostly relies on the diffusion and popularization of the scientific progress, either through direct communication with potential end-users, or by designing systems that will allow the users to feel that they keep the control on their robot.

6. Acknowledgment

This research work is partly funded by the French Institut de la société Numérique, Paris-Saclay, in the framework of the project TE2R "Traces, Explications et Responsabilités du Robot", and by the Bpifrance Romeo2 project.

7. Bibliographical References

Asaro, P. M. (2006). What should we want from a robot ethic. *International Review of Information Ethics*, 6(12):9–16.

Bertolini, A. and Palmerini, E. (2014). Regulating robotics: A challenge for europe. *Upcoming Issues of EU Law*, pages 94–129.

Capurro, R., Nagenborg, M., Capurro, R., and Nagenborg, M. (2009). *Ethics and robotics*. IOS Press.

Correa, J., McKeague, S., Liu, J., and Yang, G. (2014). A study of socially acceptable movement for assistive robots concerning personal and group workspaces. In *The Hamlyn Symposium on Medical Robotics*, page 61.

Delaborde, A. and Devillers, L. (2010). Use of nonverbal speech cues in social interaction between human and robot: emotional and interactional markers. In *Proceedings of the 3rd international workshop on Affective interaction in natural environments*, pages 75–80. ACM.

Delaborde, A. and Devillers, L. (2012). Impact of the social behaviours of the robot on the user's emotions: Importance of the task and the subject's age. In *WACAI 2012 Workshop Affect, Compagnon Artificiel, Interaction*, page 167.

Devillers, L., Tahon, M., Sehili, M. A., and Delaborde, A. (2015). Inference of human beings' emotional states from speech in human–robot interactions. *International Journal of Social Robotics*, 7(4):451–463.

Feil-Seifer, D., Skinner, K., and Matarić, M. J. (2007). Benchmarks for evaluating socially assistive robotics. *Interaction Studies*, 8(3):423–439.

Fiore, M., Khambhaita, H., Milliez, G., and Alami, R. (2015). An adaptive and proactive human-aware robot guide. In *Social Robotics*, pages 194–203. Springer.

Kasap, Z. and Magnenat-Thalmann, N. (2010). Towards episodic memory-based long-term affective interaction with a human-like robot. In *RO-MAN, 2010 IEEE*, pages 452–457. IEEE.

Lakatos, G., Gácsi, M., Tajti, F., Koay, K. L., Janiak, M., Faragó, T., Devecseri, V., Kovács, S., Tchon, K., Dautenhahn, K., et al. (2013). Dog-inspired social behaviour in robots with different embodiments. In *4th IEEE Intl' Conference on Cognitive Infocommunications., At Budapest, Hungary*.

Leary, T. (1958). Interpersonal diagnosis of personality. *American Journal of Physical Medicine & Rehabilitation*, 37(6):331.

Lin, P., Abney, K., and Bekey, G. A. (2011). *Robot ethics: the ethical and social implications of robotics*. MIT press.

Malle, B. F. (2015). Integrating robot ethics and machine

morality: the study and design of moral competence in robots. *Ethics and Information Technology*, pages 1–14.

Pointeau, G., Petit, M., and Dominey, P. F. (2013). Embodied simulation based on autobiographical memory. In *Biomimetic and Biohybrid Systems*, pages 240–250. Springer.

Riek, L. D., Hartzog, W., Howard, D. A., Moon, A., and Calo, R. (2015). The emerging policy and ethics of human robot interaction. In *HRI 2015, Portland, OR (Extended Abstracts)*, pages 247–248.

Sharkey, A. and Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology*, 14(1):27–40.

Stachowicz, D. and Kruijff, G.-J. M. (2012). Episodic-like memory for cognitive robots. *Autonomous Mental Development, IEEE Transactions on*, 4(1):1–16.

Strong, S. R., Hills, H. I., Kilmartin, C. T., DeVries, H., Lanier, K., Nelson, B. N., Strickland, D., and Meyer III, C. W. (1988). The dynamic relations among interpersonal behaviors: A test of complementarity and anticomplementarity. *Journal of Personality and Social Psychology*, 54(5):798.

Tahon, M. and Devillers, L. (2016). Towards a small set of robust acoustic features for emotion recognition: Challenges. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 24(1):16–28.

Tamburrini, G. (2009). Robot ethics: A view from the philosophy of science. *Ethics and Robotics*, pages 11–22.

van Wynsberghe, A. (2013). A method for integrating ethics into the design of robots. *Industrial Robot: An International Journal*, 40(5):433–440.

Wilson, J. R. and Scheutz, M. (2015). A model of empathy to shape trolley problem moral judgements. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 112–118. IEEE.

Multimodal Sentiment Analysis in the Wild: Ethical considerations on Data Collection, Annotation, and Exploitation

Björn Schuller[1,2], Jean-Gabriel Ganascia[3],Laurence Devillers[4,5]
[1]University of Passau, Chair of Complex & Intelligent Systems, Passau, Germany
[2]Imperial College London, Department of Computing, London, United Kingdom
[3]University Pierre & Marie Curie, LIP6, Paris, France
[4]CNRS, LIMSI, Orsay, France
[5]University Paris-Sorbonne IV, Paris, France

E-mail: [1]schuller@IEEE.org, [3]jean-gabriel.ganascia@lip6.fr, [4]devil@limsi.fr

Abstract

Some ethical issues that arise for data collection and annotation of audio-visual and general multimodal sentiment, affect, and emotion data "in the wild" are of types that have been well explored, and there are good reasons to believe that they can be handled in routine ways. They mainly involve two areas, namely research with human participants, and protection of personal data. Some other ethical issues coming with such data such as its exploitation in real-life recognition engines and evaluation in long-term usages are, however, less explored. Here, we aim to discuss both – the more "routine" aspects as well as the white spots in the literature of the field. The discussion will be guided by needs and observations as well as plans made during and for the European SEWA project to provide a showcase example.

Keywords: affective computing, sentiment analysis, ethical, legal and social implications (ELSI), data protection

1. Introduction

In the recent time we witness ever-more collection "in the wild" of individual and personal multimodal and increasing amounts of sensorial affect and sentiment data, crowd-sourced annotation by large groups of individuals with often unknown reliability and high subjectivity, and "deep" and partially less supervised learning with limited transparency of what is being learnt, and how applications depending on such data may behave. This renders the ethical, legal, and social implications (ELSI) more crucial than ever before in the field of language and multimodal resources. Accordingly, it makes the related aspects(e.g., privacy, traceability, explainability, validity, etc.) and according responsibility that comes with the collection, annotation, storing, and in particular also exploitation of (human) data of personal affect, behaviour, emotion, opinion, and sentiment a key concern. This comes in particular, as automatic systems increasingly exploit data of (and interact with) humans of all ranges (e.g., children, adults, vulnerable populations) including non-verbal and verbal data occurring in a variety of real-life contexts (e.g., at home, the hospital, on the phone, in the car, in the classroom, or within public transportation) and act as assistive and partially instructive technologies, companions, and/or commercial or even decision making systems.

In contrast to this increased relevance, the body of literature (cf., e.g., [1-5]) dealing with ELSI aspects is hardly in any balance with the number of technical publications found on the topic. Here, we aim to discuss these aspects, guided by a showcase example to provide a basis of discussion: This example will be the Automatic Sentiment Analysis in the Wild (SEWA) European project[1] that set off early in 2015. The project has the goal to advance models and algorithms for machine analysis of facial, vocal, and verbal behaviour, to realise naturalistic human-centric human-computer interaction and computer-mediated face-to-face interaction. It aims at a set of audio and visual spatiotemporal methods for automatic analysis of human spontaneous (as opposed to posed and exaggerated) patterns of behavioural cues including analysis of sentiment and liking. Technologies that can robustly and accurately analyse human facial, vocal, and verbal behaviour (and interactions) in the wild, i.e., in people's everyday life's surroundings, as observed by webcams in digital devices, would have profound impact on both, basic sciences, and the industrial sector. They could open up tremendous potential to measure behaviour indicators that heretofore resisted measurement because they were too subtle or fleeting to be measured by the human eye and ear, would effectively lead to development of the next generation of efficient, seamless opinion mining. Accordingly, one could expect profound impact on business as automatic market research analysis would become possible, and further beyond, recruitment could become more objective and green as travels would be reduced drastically at the same time, however, raising considerable ELSI implications such as whether computer-assisted recruitment is sufficiently reliable. The technology would also enable user-centric human-computer interaction by affective multimodal interfaces, and one could think of interactive multi-party games, and online services such as social TV. A large number of further applications would be enabled such as next generation healthcare technologies by remote monitoring of conditions like pain, anxiety and depression, and alike, to mention but a few examples.
This makes it obvious, what huge responsibility lies in the accuracy of such according recognition engines, their thoughtful implementation, and reasonable communication with regards to their reliability and privacy and individual rights awareness. Furthermore, learning models of human affect, behaviour, and sentiment suitable for machine analysis depends on having suitable data recordings of human behaviour to

[1] http://www.sewaproject.eu/

learn from. Hence, an important aspect of the SEWA project lies in collecting suitable datasets of sufficient labelled examples for building robust tools. Its plan includes the release of a large volume of audio-visual data of human behaviour recorded in the wild together with expert annotations in the form of a publicly available database. The intention behind is to push forward the research in multicultural and multilingual automatic human affect behavioural analysis and user-centric HCI and FF-HCI.

Here, we aim to discuss the good practices and ethical standards and issues at different stages of such collection, annotation, and exploitation of sentiment as collected "in the wild". There are two main ethical issues of concern that we will be dealing with:

- The first concerns the fact that human subjects are involved in the data collection process.
- The second concerns the use of emergent sentiment analysis technologies and their possible applications.

2. Database Collection

There are several guidelines on considerations to be made when collecting data that involves humans. Further, boards and mechanisms to overlook the process are usually in place. To give an example, in the United States of America and similarly across Europe, database collection is strongly governed by a university's institutional or ethical review board. Such a board has to approve ahead of the collection, monitor throughout the collection, and review afterwards what has been collected and potentially distributed in the context of (human) behavioural research. Similar boards are increasingly required and overlooked in connection with (inter-)national public research funding. To stick with our illustrative example, we will outline the process as encountered in the SEWA project data collection. These are given for the sake of completeness, albeit most of the outlined points are common knowledge. According to the ethical standards of human experimentation and to the requirement of the Imperial College Research Ethical Committee (ICREC), the questions relative to data collection can be subdivided into three parts:

Informed consent: A form of consent needs to best be validated by experts before the beginning of the data collection phase. This form, which will be signed by subjects involved in the experimentation, includes the data protocol description, the aim of this experiment, the description of technologies used to capture audio and visual signals and the storage and use of the data. In the SEWA project, this form needed to be translated into six languages by native speakers as the SEWA team collects multilingual and multicultural data. As the raw data will be made available to the scientific community for research purposes, the participants are made aware about the openness of the data when they sign the consent providing different levels of agreement such as usage only within the scientific community or giving consent usage of image and video material in dissemination for the scientific community or broader public. Ideally, this consent form should also explain the benefit to the public arising from the collection and the individuals taking part.

Verification of the harmless nature of the data collection. In the example of the SEWA project, there are no invasive sensors because only acoustic and video signals are used. In fact, such sensors would raise additional concerns, as it is less transparent to participants what kind of information could be contained, as they cannot access it themselves in natural ways. In addition, the video material that will be presented during the experiment is not supposed to be traumatic.

Data storage: Here, proper ethical and legal handling has to be ensured. For example, it needs to be validated whether the data are "sensitive" such as including banking information, the tax information, the health data, etc. In SEWA, the samples collected are twofold, reactions to video ads and face-to-face dyadic conversations. While the data is not sensitive in the sense as described above, anonymous storage needs to be discussed critically. Note that, anonymisation means two different things: First, to break the link between personal data and the persons whom these data are drawn from, and second, to make it impossible to retrieve the persons from their personal data. While it appears quite easy to anonymise data in the first sense, because it is sufficient to remove or to make inaccessible explicit references to the persons, i.e., their name and address, it is by far more difficult to prevent re-identification. Even when explicit references to names and addresses are removed, it appears possible by cross-references on multiple databases to infer names and addresses. Besides, the specific nature of data in the SEWA project, i.e., face images and speech signal, allows a natural re-identification by people who know the persons or using face and speaker recognition. As a consequence, if raw data are stored in the database, it is always possible to recognise people that participate to the project.

Rather than discussing the technical implications of this oversight in further detail, let us now switch to the interesting question what one cannot– or more specifically, examples of what we could not – do when collecting according data, given the named restrictions and further ethical considerations. As the desire of the SEWA project and in fact of most data collection centred around affective and behavioural computing, sentiment analysis, and opinion mining is usually to collect naturalistic data "in the wild" to be as close to the real-world use-case as possible, implications arise in particular from the multimodal nature of data: If one collects audio-visual data in the wild as is the case in SEWA, one potentially collects footage of other individuals not knowingly involved in the recording, or private information such as number plates of cars parked, the inside of private living space, etc. Thus, without ensuring massive resighting, reviewing, and processing of the collected data (such as by hiding others' faces or number plates by black bars or alike), one cannot share such recordings. However, the workload involved may be prohibitive such that, the recordings may at the end not be made in spaces outside of the property of those recorded or empty public spaces. Also, introducing such "hiding" or "blurring" elements may influence the training of machine learning algorithms.

Even more obviously, collecting data without the participants being aware of the recording may be desirable to increase the degree of spontaneity and naturalism, yet, comes at even higher ethical and legal restrictions. Very little data has been collected in the named fields in such a manner up to this point, such as in [6]. There, the local (Austrian) law allowed for private (audio-visual) recordings that may only be used for oneself unless the persons involved gave their consent (including consent given only after recording) for the material to be used for scientific (or other) purposes. In addition, as the recordings in that work took place in a private Supermarket, agreement of the shop-owner was needed in advance. As a consequence, surveillance notes needed to be put into place potentially reducing spontaneity of the behaviour. This example shows that "in the wild" collection comes at considerable efforts, but also limitations in terms of local environment and persons involved and their awareness of being recorded. Further, as the desire is often to cover for a gender, age, and cultural balance in such collection, it may be of critical relevance to decide on the material used for stimulation or induction of sentiment or affect. As a consequence, the effect may be reduced, as certain material or ways of eliciting reactions may not be appropriate to all participants of a database collection. As an example, showing extreme violence to participants may have a strong affect eliciting effect, but may not be appropriate in many cases. Similarly, some religious or political material may be sensitive to some cultural or ethnical groups in the context of sentiment analysis and opinion mining as outlined above.

A related interesting question touches upon how privacy impacts on the ability to understand the collected data. To give an example, in the SEWA project, pairs of subjects have been recorded that briefly discussed commercial spots they first watched by themselves. The precondition during enrolment for the study was that, such a pair has to know each other in advance in order to avoid (usually over-friendly and targeted towards each other rather than the subject of interest of the recordings) "getting to know" behaviour in the short time of the recording. However, owing to privacy restrictions, the full relationship status may not be known or revealed but clearly of interest when interpreting the data as to which part of behaviour shown is related to the content of the commercial or to the person being spoken to.

Further, it seems not trivial to make participants in recordings understand privacy protections leading to the question of according consequences of data collection. A first (comparably minor) "risk" is losing potential participants as they may misinterpret protection such that they refrain from participation despite the data and privacy protection mechanisms being at very high levels. In the example named above taken from [6] of subjects being recorded unknowingly at first, there is a fair chance of losing participants in a study due to their surprise of having been recorded unknowingly at first that might have agreed if they had been told in advance. Then, however, the behaviour would have likely been less spontaneous. However, a more serious risk is of the nature that subjects do not understand all implications if the privacy protection is rather weak.

To conclude this section, we provide a sketch of the collection in the SEWA project. The following balancing of participants was targeted: across age from 18 years onwards by five groups as follows: 18-29, 30-39, 40-49, 50-59, and 60-80 years. Further, as participants were grouped in pairs each knowing each other as described, the couples were balanced in terms of best even distribution per age group by female-female, male-male, and mixed gender. As a target, each age and gender constellations had to appear at least once and ideally, each should appear twice totalling up to at least 30 pairs or 60 individuals per language/culture of collection. Given the difficulty to evenly recruit according pairs across all age groups and classes, some groups such as younger individuals are present slightly more strongly in the final database.

The recording was split into two parts: In the first part, each participant had to individually watch four commercial spots with 60 seconds, each. These were chosen such as to induce different affective states including amusement, empathy, positive sentiment or boredom. A challenge at this point was to select these such as to induce target states and behaviour and at the same time not be offensive or disturbing in any way as described above. Within the second part, the participants communicated via a video-chat software to another participant known to them – on average for 4-5 minutes – regarding the content of the spots just seen each by themselves. The intention behind is to collect further reactions and opinions with respect to the content of the commercial and the product, service or charity appeal shown, which are the highlights of the spots, whether these are appropriate, how they could be improved and alike. Further, this allows for analysis of inter-human behaviour in dyadic conversations.

After obtaining ethical approval for the SEWA experiment internally and from an external ethical advisory board formed by the second and third author of this contribution, the experiment protocol was implemented and again overlooked. Next, a website and service for collection was implemented by partners of the project via which at this point 199 successful data recording sessions took place including 398 participants from the six different language and cultural backgrounds (British, Chinese, German, Greek, Hungarian, and Serbian). This required informed consent forms and the web interface to be translated into each of the six languages involved as named above. These informed them on the funding source, the intended recording and annotation in principle, the foreseen benefit to society coming from the project, and their rights to withdraw recordings at any time besides standard explanations on privacy and protection concerns. It clearly stated that, participation is voluntary, non-participation will not result in any kind of disadvantage, and that termination is possible at any moment during the recording. It also provided a contact address for independent help and information on ELSI implications at the university or responsible body. The participants had to register first on a secure web page, fill in a form of demographic questions and confirm their email address via an email sent to them. With the conductor of the experiment they then had to agree upon an appointment where both partners were

available for around 15 minutes via email. They were instructed to do the recordings at home or any other venue of choice, and that noises are no problem. However, they were not allowed to be in the same room as the partner. This freedom of choice of venue can lead to the above sketched issues of potential inclusion of other individuals or other's property which has to be counter-checked during annotation.

Each participant further used her own PC or notebook with their own webcam and microphone (intentionally) leading to a high "in the wild" variability of recording devices. They were further using their own internet connection. From an ethical point of view, this may be seen as limiting factor given potential exclusion of parts of the population. For the SEWA project, this may be less of a concern given that (most of) the use-cases address data analysis with implication of mostly such individuals that possess and use according infrastructure. This may, however, clearly be different in other studies. The recording was fulfilled via the webpage which was largely self-explanatory. Participants had to log in in time at the agreed upon time slot. Each participant had to fill in a consent form also in print version and sign and send as a scan via email. There, they had the choice to agree to usage for scientific purposes and additionally whether recordings may be used in a public context. A financial reward was given to the participants via bank transfer. Bank data could optionally be communicated via email or phone. Obviously, ethical implications also come with graduation of participants. Here, the amount was chosen small enough to be rather of symbolic nature then risking involving participants that "sell" their data.

The recordings made contain 44 hours of audio-visual footage including a wide range of spontaneous expressions of emotions and sentiment. It seems noteworthy that, due to the technical framework and requirements (higher bandwidth needed, recordings were considered only valid if successful during the first attempt, as otherwise the reactions would not be spontaneous any more) a higher rate of failure exists. To exemplify, for the recordings taken in Germany, 57 sessions were started, 43 pairs attempted, but only 37 pairs successfully recorded in the end. Obviously, this is difficult from an ethical point of view, as some participants could not be included due to technical issues, which may be disappointing to them. Further, the higher number of attempts than pairs shows repeated difficulties at the beginning prior to the recording of interest. This is time consuming for the participants and potentially influences their affect and mood. As a consequence, all efforts were made to avoid such circumstances. There was no pronounced gender effect for within-gender and cross-gender pair differences (i.e., all three constellations of gender grouping occurred equally often).

The collection further included extraction of acoustic, linguistic, and visual features. Acoustic features were extracted in two different sizes of feature space by the open-source openSMILE [7] ComParE and GeMAPSv01a standardised feature sets from all SEWA recordings. Similarly, 49 facial landmarks were automatically tracked. This provides an interesting alternative option of distribution of data: Rather than distributing the full audio-visual recordings, sharing just feature representations for reproduction and comparison of and with scientific findings comes at higher protection of privacy. However, care is needed, as features sampled at short intervals and in high numbers and complementarity may allow for resynthesis of the original (audio-visual) source data to large extents. Further compression such as by (sub-band) vector quantisation may reduce this risk [8].

3. Data Annotation and Release

Data annotation bears its own ELSI pitfalls in particular in the context of crowdsourcing given that, the data will be shown to potentially unknown raters "outside the lab". This may include them watching the material in public spaces in the presence of others, as crowdsourcing increasingly becomes mobile (cf., e.g., [9]). In [10], the authors name the primary concern of 12 researchers questioned in an according study to be privacy-related ranked second after accuracy-related and further concerns such as related to the reliability and the costs involved when it comes to crowdsourced video coding. The authors further suggest blur filters as suited means to better hide the identity of the individuals to be coded. Unfortunately, as one may expect, this does at the same time downgrade also the coders "ability to accurately and reliably code behaviours" [10]. Luckily, however, the decrease was "not as steeply as the identity test"[10]. Accordingly, such methods need to be improved, and similar methods need to be established and evaluated carefully for audio or even textual and further information "blurring" in this context.

For SEWA data annotation, annotation within the lab including the crowdsourcing platform iHEARu-PLAY [11] was successfully used up to this point. The latter provides a gamified approach without monetary compensation. Specific ethical considerations are summarised in detail in [12]. The scheme for the annotation of the data includes continuous assessment along the three primitives or dimensions arousal, valence, and sentiment/liking. A major issue in this respect is to correctly instruct annotators such as to ensure good understanding of the differences between these primitives to warrant high quality annotations. Further, verbal transcriptions including non-verbal vocalisations were made manually and counter-checked in five languages up to this point (excluding Greek). Here, it was necessary to identify native speakers of these languages, each, for the transcriptions to ensure accurate transcription. The results are overall further refined through semi-automatic correction.

A core SEWA dataset (currently 540 representative segments – 90 from each culture group – chosen in balance by high/low arousal, high/low valence, and liking/disliking) is currently further annotated fully in terms of facial landmarks, facial action units (FAUs), mimicry, sentiment, rapport, and template behaviours. Again, this will partially require expert coders – in particular Facial Action Coding System (FACS) certified coders for the FAUs. This shows that only part of the annotation can be distributed to the (partially laymen) crowd. To ensure high privacy standards given the "in the wild" nature of the collection, the SEWA database shall be

used at first within the project consortium to identify potential remaining issues internally during application and use of the data prior to a public release. This release will be via a web-portal allowing for enhanced search functionality to invite also non-technical scientific usage where fast retrieval of specific behaviour is crucial. In fact, we believe a broad release for research only (e.g., by password protected access via secure sites) or potentially even the greater public to be of crucial importance: First of all, giving access to other researchers will avoid double efforts in collection and thus require less participants and annotators. At the same time, this can accelerate progress that may be highly needed such as in the health sector. Further, it increases reproducibility of findings – an often violated key principle of good research. This can be done, e.g., in competitive challenges to increase interest in the data such as in [13] or the MediaEval series. Finally, the data collection often is subsidised by grant money from public sources – thus, the public should best benefit from the efforts and resources should be spent in an utmost efficient manner, only.

4. Exploitation

Data collected and annotated in the context of affective and behavioural computing and sentiment analysis is usually used to train models for applications including analysis and synthesis of emotion, sentiment, and behaviour. In the SEWA project, application of recognition of human sentiment and behaviour includes in particular recommendation systems and face-to-face interaction through a chat roulette social game. In these applications, the data storage is a challenging topic. The system architecture solution proposes local data storage to protect privacy. Similarly important are, however, ethical implications of the actual application. In the project, two focus groups were built to ensure responsible and sensitive discussion: the first includes the Ethical Advisory Board (EAB) of the project – as outlined above, instantiated by the second and third author of this contribution, and members of and industrialValorisation Advisory Board (VAB); the second comprises users and professionals. The following key points are considered of interest by these boards and in discussions:

Recognition, recognisability, and uncertainty: It needs to be ensured that what is being recognised by an automatic system is recognisable at all. One easily falls for the trap of taking it for granted that computerised measurement and classification are objective, as they stem from a technical system. They thus would lead to formalised representation of human emotion and disposition. However, many human phenomena including the above are too complex and ambiguous to allow for (complete) objectification. This comes among others, as higher level individual aspects and context need to be taken into account, but often are not. Proper communication of the recognisable thus is of crucial importance, such as by provision of confidence measures and implementation of benchmark tests such as the Interspeech Computational Paralinguistic Challenges 2009-16 or the Audio/Visual Emotion Challenges 2011-16 (cf. e.g., [13]). Further, the uncertainty has to be protected, i.e., it has to be ensured that certain private spheres are not entered and users of technology are aware of a remaining uncertainty.

Reductionism: Models designed for computational assessment of human emotion, sentiment, and behaviour are often simplified. This bears the danger of unforeseeable implications as the actual problem's complexity is reduced to a potentially insufficient representation.

Effect of erroneous decisions: The harmless character of erroneous decisions has to be ensured in best possible ways. In a recommender system such as envisioned by the SEWA project as one exemplary use-case, the implications may be less severe such as receiving sub-optimal recommendations on the content of potential interest, e.g., music or movies. However, in the second use-case of a social chat-roulette game, implications are more severe: If a system makes wrong assumptions on users (dis-)liking each other, the social implications may be (more) drastic such as (erroneously) made to belief someone dislikes the other. Clearly, however, there are potentially even more critical use-cases such as the above named "green" job interviews where a system may become responsible of someone erroneously not being employed.

As industrial partners and health and security providers increasingly collaborate with scientists rooted in computer science, and electrical engineering in the fields of affective computing, sentiment analysis, opinion mining, behavioural and social signal processing, it is increasingly important to understand what can or cannot be modelled and sensed in an accurate and reliable fashion. It will be important to also further strengthen the collaborative and communicative aspect in this respect.

5. Conclusion

Many ethical issues (evaluation of the sentiment analysis technologies in the wild, possible applications, etc.) need to be addressed when dealing with affective and behavioural corpus collection, annotation, and exploitation. Here, we named key-aspects, and exemplified them in the context of an ongoing European project dealing with "in the wild" collection across six cultures / languages. The idea was to demonstrate by a case study how broader ethical principles can be translated into a concrete policy. However, additional experience can be expected throughout the further runtime of the SEWA project contributing to its detailed policy to be shared. Future implications may be even more challenging, once technical systems become increasingly "conscious" also in emotional ways [14-18].

Acknowledgements

The research leading to these results has received funding from the European Union's Horizon 2020 Programme research and innovation programme under grant agreements Nos. 645094 (SEWA) and 644632 (MixedEmotions), and from the German national BMBF IKT2020-Grant under grant agreement No. 16SV7213 (EmotAsS). The content presented has benefited from discussion with Harald Traue of University of Ulm and further participants of the ELSI networking workshop InterEmotio organised by the German BMBF in Stuttgart/Germany on 26 February 2016. We also thank the members of the SEWA consortium and its industrial Valorisation

Advisory Board. The authors further acknowledge the three (anonymous) reviewers' highly productive suggestions.

Bibliographical References

[1] Reynolds, Carson, and Rosalind Picard. "Affective sensors, privacy, and ethical contracts." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2004 Extended Abstracts on Human Factors in Computing Systems, pages 1103-1106, Vienna, Austria, ACM, 2004.

[2] Reynolds, Carson Jonathan. "Adversarial uses of affective computing and ethical implications." Dissertation. Massachusetts Institute of Technology, 2005.

[3] Goldie, Peter, Sabine Döring, and Roddy Cowie. "The ethical distinctiveness of Emotion-Oriented Technology: Four long-term issues." Emotion-Oriented Systems. Springer Berlin Heidelberg, 2011. 725-733.

[4] Duffy, Brian R. "Fundamental issues in affective intelligent social machines." Open Artificial Intelligence Journal 2.1 (2008): 21-34.

[5] King, Elizabeth W. "The Ethics of Mining for Metadata Outside of Formal Discovery" Penn State Law Review 113 (2008): 801.

[6] Eyben, Florian, Weninger, Felix, Paletta, Lucas, and Schuller, Björn. "The acoustics of eye contact – Detecting visual attention from conversational audio cues" Proceedings 6th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Gaze in Multimodal Interaction, GAZEIN 2013, held in conjunction with the 15th International Conference on Multimodal Interaction, ICMI 2013, pages 7–12, Sydney, Australia, ACM, December 2013.

[7] Eyben, Florian, Weninger, Felix, Groß, Florian, and Schuller, Björn. "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor" Proceedings of the 21st ACM International Conference on Multimedia, MM 2013, pages 835–838, Barcelona, Spain, ACM, 2013.

[8] Zhang, Zixing, Coutinho, Eduardo, Deng, Jun, and Schuller, Björn. "Distributing Recognition in Computational Paralinguistics." IEEE Transactions on Affective Computing, 5.4:406–417, October–December 2014.

[9] Nebeling, Michael, To, Alexandra, Guo, Anhong, de Freitas, Adrian A., Teevan, Jame, Dow, Steven P., and Bigham, Jeffrey P. "WearWrite: Crowd-Assisted Writing from Smartwatches" Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2016, 10 pages, San Jose, CA, ACM, 2016.

[10] Lasecki, Walter S., Gordon, Mitchell, Leung, Winnie, Lim, Ellen, Bigham, Jeffrey P., Dow, Steven P. "Exploring Privacy and Accuracy Trade-Offs in Crowdsourced Behavioral Video Coding" Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2015, 6 pages, Seoul, Republic of Korea, ACM, 2015.

[11] Hantke,Simone, Eyben, Florian, Appel, Tobias, and Schuller, Björn. "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing." Proceedings 1st International Workshop on Automatic Sentiment Analysis in the Wild, WASA 2015, held in conjunction with the 6th biannual Conference on Affective Computing and Intelligent Interaction, ACII 2015, pages 891–897, Xi'an, P. R. China, AAAC, IEEE, September2015.

[12]Hantke, Simone, Batliner, Anton, and Schuller, Björn."Ethics for Crowdsourced Corpus Collection, Data Annotation and its Application in the Web-based Game iHEARu-PLAY" Proceedings of the1st International Workshop on ETHics In Corpus Collection, Annotation and Application, ETHI-CA² 2016, satellite of the10th Language Resources and Evaluation Conference, LREC2016, 6 pages, Portoroz, Slovenia, ELRA, May 2016.

[13] Ringeval, Fabien, Schuller, Björn, Valstar, Michel, Jaiswal, Shashank, Marchi, Erik, Lalanne, Denis, Cowie, Roddy, and Pantic, Maja."AV+EC 2015 - The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data." Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC15, collocated with the 23rd ACM International Conference on Multimedia, MM 2015, pages 3–8, Brisbane, Australia, ACM, October 2015.

[14] Sullins, John P. "Robots, love, and sex: The ethics of building a love machine." IEEE Transactions on Affective Computing, IEEE 3.4 (2012): 398-409.

[15] Torrance, Steve. "Ethics and consciousness in artificial agents." AI& Society 22.4 (2008): 495-521.

[16] Whitby, Blay. "Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents." Interacting with Computers 20.3 (2008): 326-333.

[17] Pontier, Matthijs A., and Johan F. Hoorn. "Toward machines that behave ethically better than humans do." Proceedings of the 34th international annual conference of the cognitive science society, Cognitive Science 12 (2012).

[18] Grinbaum,Alexeij, Chatila, Raja, Devillers, Laurence, Ganascia, Jean-Gabriel, Tessier, Catherine, Dauchet, Max. "Ethics in Robotics Research" Proceedings International Conference on Robot Ethics, ICRE 2015, Lisbon Portugal, October 2015.

Subscribing to the Belmont Report: The Case of Creating Emotion Corpora

Jocelynn Cu*, Merlin Teodosia Suarez* and Madelene Sta. Maria^
*Center for Human Computing Innovations, De La Salle University
^ Research Ethics Office, De La Salle University
2401 Taft Ave., Manila, Philippines
E-mail: jocelynn.cu@dlsu.edu.ph, merlin.suarez@dlsu.edu.ph, madelene.stamaria@dlsu.edu.ph

ABSTRACT

Works on human emotion and behavior analysis has been on the increase in recent years. This is primarily the result of maturity of information technology, evidence that it has enmeshed itself in a human's everyday activities. Current approaches to emotion and behavior modeling require the creation of corpora from human subjects typically engaged in interactions. Collection of information and other data from humans necessitates following certain guidelines to ensure their privacy, security and over-all well-being. This work presents how the Belmont Report may be adapted in the practice and review of acceptable standards in the creation of emotion corpora for developing countries, given that these communities do not possess high awareness of their privacy rights. It describes the creation of several multi-modal corpora of patients and students, including the ethical practices employed following the principles indicated in the Belmont Report. At the end of the paper, recommendations how to better improve research practices are shared, including possible research directions.

Keywords: emotion corpus, research ethics, human subjects

1. Introduction

The creation of emotion corpora is one of the most important tasks a researcher of affective computing has to do. Most emotion models are created by extracting patterns over an emotion corpus using machine learning techniques and generalizing over these. Using these approaches, the quality of the models is heavily reliant on the quality of the corpus used. While some emotion corpora were created by analysing TV shows and movies, some researchers prefer naturalistic, multimodal, and annotated with multiple levels of label. Data created in this way involves seeking the involvement of human subjects, designing interaction parameters, and then recording emotional expressions. Typically, audio-visual information is recorded. Some works also collect physiological information collected from sensors such as EEG, respiration, and skin conductance. Creation of these such corpora help increase the utility of the emotion models when deployed in real-world scenarios.

Questions of ethical concern confronting researchers building emotion and affect corpora are those related to the privacy of participants and the anonymity of data. These issues become most important when emotional expressions are sampled in public places (Liikkanen et al, 2009), when the subjects of natural emotions may oppose the circulation of data in the community (Fu et al, 2012), or when participants are made to respond in situations that make them exhibit strong or unpleasant emotions (Bänziger et all, 2006; Vidrascu & Devillers, 2006). Concerns about privacy and anonymity are likewise important when data are collected in real-life interactions (Lubis et al, 2015; Marioryad et al, 2014; Busso & Narayanan, 2008).

Researches in the field are regulated by existing ethical practices enforced by institutional review boards. Informed consent given to participants is a usual ethical practice expected by most boards (Hill et al., 2013; Fu et al, 2012; Douglas-Cowie et al., 2007). Constraints in giving out participant demographic information (Hill et al., 2013), or providing vague and nonspecific information about actors displaying information are prescribed by review boards to prevent easy identification, or to ensure confidentiality. The non-diffusion of the corpus to the community also serves to preserve the privacy of data (Vidrascu & Devillers, 2006). However, problems in maintaining participant anonymity continue to challenge researchers in the field. For example, a real difficulty would include the possibility of anonymizing data without losing essential content from facial videos (Liikkanen et al, 2009).

Most researchers are aware of practices that are clearly unethical and would do away with these in researches. These include recording unobtrusively by concealing equipment when recording participants (Campbell, 2006), or inducing strong emotions in a laboratory setting (Pelachaud, 2013). The ethical issues have mostly focused on procedures in the collection of data, i.e., the research process, and not enough concern is given to the content of research. Ikonen et al (2009) makes this distinction when they pointed out that research ethics practice addressing process is based on established medical context-based ethics reviews. Researchers building technology through the building of emotion corpora should likewise take into consideration research content issues, i.e., the consequences of the development and use of these corpora.

It is a given that the well-being and welfare of human subjects who participate in research should be considered. However, we believe that human subjects from developing countries require additional considerations. In particular, the awareness and practice of privacy protection is not as popular in developing countries (Hosein, 2011) mainly because privacy is perceived as a technology-based/ industrialization issue. The London School of Economics reported that most e-health systems deployed in developing countries severely lacked the privacy and security aspects, and that local awareness of privacy responsibilities is needed to make such efforts successful (LSE, 2010).

To address these concerns, we used the Belmont Report to guide us in considering ethical practices for collecting data from human subjects. It was drafted by the National Commission for Protection of Human Subjects of Biomedical and Behavioral Research, as commissioned by the United States Congress (Rice, 2008). The report presents the ethical principles and guidelines of *respect for persons*, *beneficence*, and *justice*, to protect the human subjects in the conduct of research.

In this paper we present practices employed at De La Salle University in the Philippines in the last five years to create various kinds of databases used for emotion and behaviour modelling as indicated by the Belmont Report. These databases were chosen to show the variety of participants (from children to adults), data collected (everyday activities, physiological, and audio-visual information), and domain (health and wellness and education). In particular, various attributes were extracted, from modalities (visual, audio, body movement and gestures, and physiological data) using obtrusive (i.e. Emotiv Epoc) to non-intrusive devices (video camera), to typical user profile information like age, gender, anxiety score, among others. Emotional responses were either induced or spontaneous, i.e., recorded as they occur during an interaction with another person or while performing an activity (such as drawing, or answering a Math exercise).

2. Emotion Corpora Creation for Innovative Human-Machine Interfaces

Several data sets for emotion and behavior analysis (Table 1) were created by the Center for Human Computing Innovations (CeHCI) over the years for various research projects related to developing socially-intelligent human-machine interfaces.

The datasets were primarily created to build emotion models from social signals such as laughter, and for applications such as wellness and health, and education. Multimodal data was collected while human subjects were engaged in human-human interactions, or as they were engaged in a particular activity/task.

The work of Chuacokiong and Suarez (2012) created a database of a person's daily activities annotated with emotions as he occupied a sensor-rich space. The subject annotated his activities with arousal and valence values as he performed these. This data was used to make predictions about his future emotions and provide pro-active support.

A spontaneous laughter corpora was created (Imperial and Cu, 2015; Luz et al, 2015) to analyze the emotion carried during human-human interactions. Participants (aged 8 – 13, and 18-24 years old, respectively) watched funny video clips, joked with peers, and talked about their personal experiences were video-recorded. The corpora were used to analyze audio characteristics of children laughing, and investigated the dynamics of the body during laughter occurrences.

Lim and Suarez (2015) studied the effects of odorants on human stress levels. To do this, they created a database of its subjects' physiological signals, specifically blood volume pulse, respiration and skin conductance data as they were answering the modified Stroop Test, meant to induce stress in the participants. The experiment was conducted in two adjacent rooms, three times per subject. The first run was to answer the test without any odorant introduced into the room. The second run was held in an adjacent room which has been exposed to lavender. The third run was conducted in the original room which has been exposed to bergamot. The subject answered the Stroop Test which was modified for each run. Faculty members aged 21 to 50 years old participated in the study.

Works related to developing intelligent tutors involved studying student's affect to provide relevant remediation in the course of a learning session. The works of De Los Reyes et al (2013) and Swansi et al (2015) created a database of students' academic affective states as they learned new concepts, i.e. learning English as a second language and acquiring computer literacy skills, respectively. De Los Reyes et al (2013) recorded learning sessions between a human teacher and pre-school students learning English. It was annotated with academic emotions, namely boredom, confusion, delight, engagement, frustration, and neutral.

The work of Swansi et al (2015), on the other hand, focused on recognizing academic emotions of adult learners using the theory of Androgeny. Electroencephalogram (EEG) signals were taken from subjects to determine their emotional state while solving exercises.

Arce et al (2014) and Calpo et al (2014) collected physiological data of children with autism, focusing on their arousal levels as they listened passively to music and answered a Math exercise, or created drawings. The children were aged 7 to 12 years old, and their skin conductance were collected the entire time data collection session.

Authors	Purpose of the datasets	Types of data gathered	Label	Type of Expression	Emotion Labels	Type of Subject and Age	Disclosure of Research Objectives
Chuacokiong and Suarez (2012)	Predicts the person's activity inside the empathic space	activities (e.g., studying, resting/sleeping, using a computer, etc.)	Dimensional	Spontaneous	arousal, valence	4 adults (18 ~ 20 yo)	full disclosure
De Los Reyes et al (2013)	Identifies the academic affective state of the child while learning English words	facial expression, speech	Categorical	Spontaneous	academic affective states (boredom, confusion, delight, engagement, frustration, neutral	21 children (4 ~ 6 yo)	full disclosure, waiver signed by parents
Arce et al (2014) Calpo et al (2014)	Measures the arousal level of children with autism while listening to music and doing art sessions/math problems	biosignals (skin conductance), video/art session or math experience (45-60 mins)	Dimensional	Spontaneous	personal models	6 children (7 ~ 12 yo)	voluntary, with letter of consent
Imperial and Cu (2015)	Identifies the age, gender, and affect of the person laughing	laughter, age, gender	Categorical	Induced	kinikilig (giddy), nasasabik (excitement), nahihiya (embarrassment), natutuwa (happiness), mapanakit (hurtful)	10 children (8 ~ 13 yo)	full disclosure, waiver signed by parents
Luz et al (2015)	Identifies the affect in the laughter based on body movement	body movement (head, shoulders, hand, body), laughter	Categorical	Spontaneous	kinikilig (giddy), nasasabik (excitement), nahihiya (embarrassment), natutuwa (happiness), mapanakit (hurtful)	9 adults (18 ~ 24 yo)	full disclosure
Lim and Suarez (2015)	Measures the stress level of a person in the presence or absence of odorants via physiological signals	biosignals (blood volume pulse, respiration, skin conductance, Beck Anxiety Inventory), age, work position, gender, # of hours spent on work	Categorical	Induced	personal models	adults (21-50 yo)	voluntary, with letter of consent
Swansi et al (2015)	Identifies the academic affective state of the adult while learning to use a computer	biosignals (EEG)	Categorical	Spontaneous	engaged, confused	4 adults (> 45 yo)	full disclosure

Table 1. Summary of datasets developed at CeHCI

3. Ethical Considerations in Collecting Data from Human Subjects: The Belmont Report

In an experiment conducted from 1932 to 1972, hundreds of Black patients from Alabama, with syphilis, were observed, but were untreated. Further, the human subjects were not informed of, nor consented to, their research participation (Silver, 1988). Along with other research misconducts from 1950 to 1974, this infamous experiment, called the "Tuskegee Study," prompted the United States Congress to pass the National Research Act of 1974 which created the National Commission for Protection of Human Subjects of Biomedical and Behavioral Research (Rice, 2008). The National Commission, composed of experts on ethics, religions, law, industry, and medicine, met several times from 1975 to 1978, to deliberate on the complexity of ethical problems in research (Rice, 2008). In 1978, the National Commission issued the Belmont Report[1], which describes the three fundamental principles for ethical human subjects research: respect for persons, beneficence and justice (Rice, 2008).

These principles are put into practice in research through the implementation of the informed consent procedure, a risk-benefit assessment and the selection of subjects

[1] http://www.hhs. gov/ohrp/humansubjects/guidance/belmont.htm

(Irving, 2013). Up until now the ethical considerations detailed in the Belmont report are used to ensure the protection of human subjects in research (Zucker, 2013).

3.1 Respect for Persons

The principle of respect requires researchers to preserve the subjects' right to self-determination, to treat subjects as agents of autonomy, and to provide additional protection for those who have diminished autonomy (Irving, 2013; Rice, 2008).

In our studies that involved adult subjects with ages 18 and up (e.g., Lim and Suarez, 2015; Luz et al, 2015; Swansi et al, 2015; Chuacokiong and Suarez, 2012), participation is voluntary and full disclosure is given prior to the start of the experiment. Letter of consent are signed by the subjects themselves giving the researcher full control over their data. Participants are given "tokens of appreciation" in the form of a small honorarium or free snacks at the end of the experiment. For students who are encouraged by the teacher to participate, no extra credits were given.

Subjects such as children with ages 4 to 13 years old (e.g., Imperial and Cu, 2015; Calpo et al, 2014; Arce et al, 2014; De Los Reyes et al, 2013) are considered individuals with diminished autonomy. In the case of Calpo et al (2014) and Arce et al (2013), the subjects are children diagnosed in the Autism Spectrum. In these instances, consent to participate in the experiment was given by the parent or guardian. Full disclosure was given to the parent prior to signing the consent forms. All information related to the data collection, including its procedure, devices, dangers and hazards, and benefits to the participant are discussed. Participation is voluntary. They were given tokens such as pencils, crayons, chocolate bars or candies to show the researchers' appreciation after the experiment session. Parents do not receive any form of token for allowing their children to participate.

In developing countries, it is the moral obligation of the researchers to protect their human subjects. While the Belmont Report typically considers adults as autonomous agents, the lack of sensitivity to privacy in developing countries should lead researchers to consider them as individuals with diminished autonomy.

3.2 Beneficence

The principle of beneficence requires that the research should maximize benefit and minimize harm, and should ensure the subjects' well-being (Irving, 2013; Rice, 2008).

Generally, subjects are not forced to participate in the experiment. To ensure their anonymity, their names and other personal information not relevant to the research are withheld from the final database; in cases when facial features are not used, faces were blurred in the final database and in the publications.

The objectives of the experiment including all proponents' identities are disclosed to the participants. In Swansi et al (2015), all human subjects are given the right to withhold some data if he wishes, and these are deleted immediately.

In studies that involve children, those who refuse to participate even after consent was given by the parent are not forced to participate. In Imperial and Cu (2015), data collection was done in the field, i.e., outside the University campus area. The researchers set up a mobile lab in a residential area. Parents and guardians are allowed to observe the data collection session if they wish. On the other hand, in De Los Reyes et al (2013), the children are fetched to and from their pre-school by the researchers to ensure their safety. Parents are allowed to accompany the children to the lab if they so wish, but not allowed inside the experiment area. In the case of Arce et al (2013) and Calpo et al (2014), only those children who are diagnosed to be free from sensory sensitivity are chosen to participate. This is to ensure that they will not suffer from any kind of discomfort while wearing the Affectiva Q-sensor.

It should be noted that the kinds of devices employed matter in risk-benefit analysis because these influence the methodology heavily. For example, if the Q-sensor had an alternative wrist-band, the study of Are et al (2013) would have been able to include other children with sensory sensitivity issues. Advancements in sensors and data collection devices will make similar studies more inclusive.

3.3 Justice

The principle of justice relates to the distribution of risk, such that the protection from systematic exclusion of persons is guaranteed, especially for those who will directly benefit from the research (Rice, 2008). The principle of justice demands that the provision of advantages should not only be given to those who can afford them, but also to those who are the potential users of the subsequent research applications (Irving, 2013).

For most of the data sets, other members of the community are the ones who will benefit from the technology in the future. However, in specific cases, such as in Swansi et al (2015), the adult learners get to look at the readings of the EMOTIV Epoc and the researcher explains the results. The subjects can then evaluate and reflect on the learning session and suggestions were given to the participants to help them improve their performance. In Lim and Suarez (2015), the faculty members were relaxed after undergoing the experiment due to exposure to odorants designed to relieve them of stress. For the children who participated in the experiment (De Los Reyes et al, 2013), they received free additional sessions on learning English.

Although existing measures to protect the subjects in the form of informed consent and non-diffusion of corpus to community are in place, additional measures are needed in developing countries like the Philippines. For instance, ethical policies should require researchers to educate human subjects regarding their rights to privacy and security, discuss implications and consequences of their

personal experiences and data when shared. The notion of anonymity and confidentiality in research activities should also be stressed. If the subject is a minor, the responsibility falls on the parent, guardian or the researchers.

Researchers should be required to monitor the well-being of the subjects after data collection in cases when they were subjected to a strong emotional experience (such as collecting data from horror game players). Researchers should be educated that the subject should not suffer in any way, physically or psychologically, when participating in the experiments.

For research laboratories, policies and guidelines should be set for ethical practices, sharing of databases, and dissemination of experiment results.

4 Observations

Certain observations were noted while collecting and preparing these data in relation to ethical practices. These were related to the participants' attitude towards participation and data collection, and the effect of advancement in sensors on the design of data collection methodologies.

4.1 Culture, Norms and Practices

It was observed that compared to developed countries, human subjects in the Philippines are not conscious about their privacy, particularly about who has access to their data, consent to its use and its security. Subjects do not seek ethical clearances or formal letter of request from researchers, as long as it was approved by the organization they belong to or any higher authority. They are not familiar with non-disclosure and confidentiality agreements, and therefore tend to participate in the study even in its absence.

The notion of privacy and personal information is different depending on the culture of the researchers (as those who craft the letters of consent and non-disclosure agreements, for instance) and the human subjects (as those whose data will be collected, and made to wear sensors or be exposed to cameras and microphones). The relationship between the researcher and the subjects dictates the willingness of the subject to participate in the data gathering work. If the researcher is part of the human subject's in-group, the latter is more likely to provide personal information because of the trust that exist between them.

There are two differing cases in relation to giving tokens, financial rewards, or academic incentives. A typical practice is to encourage students to participate in studies by giving them extra credits in their courses. This attracts two types of two students: those who need the extra credit to pass the course, and those who want the extra credit to receive academic honors. On the other hand, relatives and friends tend to help people within their social circle by participating in studies freely, without any token or financial reward. Either way brings about issues on biases in terms of selection of subjects.

4.2 Data Gathering Devices

The creation of corpora relies heavily on the kind of and availability of data gathering devices. These range from various kinds of cameras and microphones to wearable and physiological sensors. The methodology is dependent on the devices that are available to the researchers. For instance, wireless physiological sensors are less intrusive during data collection and are favorable to human subjects. In another case, the cost of motion-capture sensors is prohibitive, and Kinect sensors are an alternative. It was observed that the methodologies need to be adjusted depending on the subjects as well. For instance, speakers are a good alternative to earphones when children with autism were asked to listen to music.

The degree of intrusion (and possibly harm) relies heavily on the types of sensors as well. The use of brain-computer interface devices has caused discomfort to human subjects when worn for more than 20 minutes. The use of saline solution on the probes has also been an issue for some. The arrival of newer headsets, and the availability of dry probes of EEG devices will help improve the level of convenience during data collection.

5 Concluding Remarks

This paper presents how the three principles of respect for persons, beneficence and justice indicated in the Belmont Report was adapted in the creation of various emotion corpora particularly for human subjects from developing countries. The Belmont Report outlined these principles that fit the needs of researchers dealing with human subjects and behavior analysis. The practice of informed consent, risk-benefit analysis and unbiased subject selection should be continued. Some recommendations are provided below to improve researchers' and human subjects' awareness of the conduct of ethical research.

Based on our experience building these databases and in relation to the Belmont Report, it should be stressed that human subjects from developing countries be considered as individuals with diminished autonomy, as their circumstances lead them to be easily manipulated. The responsibility to educate their subjects about their rights, and the principles of respect for persons, beneficence and justice should rest on the researchers. Universities and research laboratories in developing countries should also be pro-active in training its researchers and academic staff on the conduct of ethical research.

Furthermore, researchers should be reminded to select devices that pose the lowest risk, and are non-intrusive. These devices determine the level of exposure of human subjects to any kind of harm during the data collection process, and should therefore be given attention during the ethical review.

In consideration of the principle of beneficence, careful

studies should be made on the effect of these databases in human-machine interaction. For instance, most databases are utilized to analyze, recognize, and synthesize human emotions and social signals. While impact and utility studies have been conducted, i.e. how useful is it for a child learning English as a second language to interact with an embodied conversational agent, studies as to how these affect a subject's concept of self and personhood needs to be conducted, i.e. will the child develop a healthy concept of self when it continually "learns" from a machine? What is the effect of a negative comment from such a machine? Will the child establish a healthy "relationship" with its software teacher? It should be good to initiate studies of this nature, as we prepare to deploy emotion-intelligent systems in the wild.

6 Bibliographical References

Arce, M. E., Limson, M., Ormoc, M., Yap, L., and Suarez, M. T. (2014). *Creating affect models of autism individuals using music features and physiological readings.* In the Proceedings of the 14th Philippine Computing Science Congress. March 6-8, 2014, Davao City, Philippines. Manila: Computing Society of the Philippines.

Bänziger, T., Pirker, H. & Scherer, K. (2006). *GEMEP – Geneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions*. In L. Deviller et al (Ed), Proceedings of LREC 2006 Workshop on Corpora for Research on Emotion and Affect, pp. 15-019, Genoa, Italy.

Busso, C., & Narayanan, S. S. (2008). *Scripted dialogs versus improvisation: lessons learned about emotional elicitation techniques from the IEMOCAP database*. In Proc. INTERSPEECH, pp. 1670-1673, Brisbane, Australia.

Calpo, J.J., Subido, J. E., and Suarez, M.T. (2014). *CAGeS-CB: Children with autism gesture recognition and stress level corpus*. Unpublished undergraduate thesis. De La Salle University, Manila, Philippines.

Campbell, N. (2006). *A language-resources approach to emotion: corpora for the analysis of expressive speech*. In Proceedings of LREC 2006 Workshop on Corpora for Research on Emotion and Affect, Genoa, Italy.

Chuacokiong, M., and Suarez, M. T. (2012). *Applying appraisal to detect emotions in a real-world, multi-tasking empathic space.* Unpublished masters thesis. De La Salle University, Manila, Philippines.

De Los Reyes, J. E. A., Garcia, J. G., Santiago, R. C. C., and Talento, M. E. P. (2013). *An empathic embodied conversational agent for learning english as a second language.* Unpublished undergraduate thesis. De La Salle University, Manila, Philippines.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M.. & Amir, N. (2007). The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. In *Affective computing and intelligent interaction*, pp. 488-500, Springer Berlin Heidelberg.

Fu, L., Jin, H., & Wu, X. (2012). *Design and enhancement of mandarin emotional speech database*. Lecture Notes in Information Technology, Volume 10, 212.

Hill E, Han D, Dumouchel P, Dehak N, Quatieri T, et al. (2013) *Long term SuboxoneTM Emotional Reactivity as measured by automatic detection in speech.* PLoS ONE 8(7): e69043. doi:10.1371/journal.pone.0069043.

Ikonen, V., Kanerva, M., & Kouri, P. (2009). Emerging Technologies.

Imperial, M. Z. C. and Cu, J. (2015). *Analysis of affective laughter using music-based acoustic features.* Unpublished masters thesis. De La Salle University, Manila, Philippines.

Irving, D. N. (2013). *Need to know: Nuremburg Code, Declaration of Helsinki, Belmont Report, OHRP*. Available at http://www.lifeissues.net/writers/irv/irv_214needtoknow.html.

Liikkanen, L. A., Jacucci, G., & Helin, M. (2009, September). *ElectroEmotion - A tool for producing emotional corpora collaboratively.* In Affective Computing and Intelligent Interaction and Workshops, 2009. 3rd International Conference on, pp. 1-7.

Lim, J. R., and Suarez, M. T. (2015). *Modelling Physiological Effects of Odorants in the Same Chemical Class on Stress Levels of College Faculty Members.* Unpublished masters thesis. De La Salle University, Manila, Philippines.

LSE (2010). *Electronic Health Privacy and Security in Developing Countries and Humanitarian Operations,* London School of Economics: Policy Engagement for the International Development Research Centre.

Lubis, N., Sakti, S., Neubig, G., Toda, T., & Nakamura, S. (2015). *Construction and analysis of social-affective interaction corpus in English and Indonesian*. In Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2015 International Conference, pp. 202-206.

Luz, M. B., Nocum, M., Purganan, T. J., Wong, W. S., and Cu, J. (2015). *Automatic recognition of affective laughter from body movements*. In 2015 DLSU Research Congress, 4-5 March 2015, Manila.

Marioryad, S., Lotfian, R., & Busso, C. (2014). *Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora.* In Proc. INTERSPEECH, pp. 238-242.

Pelachaud, C. (2013). *Emotion-oriented systems*. John Wiley & Sons.

Rice, T. W. (2008). *The historical, ethical, and legal background of human-subjects research.* Respiratory care, 53 (10), pp. 1325-1329.

Silver, G. A. (1988). *The infamous Tuskegee Study*. American Journal of Public Health, 78 (11), 1500.

Swansi, V., Cu, J., Azcarraga, J., and Suarez, M. T. (2015). *Investigating academic affective states of Sub-Saharan African, Filipino and Latin American Spanish adult learners using brainwave signals.* In 6th Int'l Workshop on Empathic Computing, 26-30

October 2016, Italy.

United States. (1978). *The Belmont Report: ethical principles and guidelines for the protection of human subjects of research.* Bethesda, Md.:The Commission.

Vidrascu, L. & Devillers, L. (2006). *Anger detection performances based on prosodic and acoustic cues in several corpora*. In Workshop on Corpora for Research on Emotion and Affect, pp. 13-16.

Vidrascu, L., & Devillers, L. (2006). *Real-life emotions in naturalistic data recorded in a medical call center.* In Proceedings of LREC 2006 Workshop on Corpora for Research on Emotion and Affect, pp. 20-24, Genova, Italy.

Zucker, D. (2013). *The Belmont Report*. Encyclopedia of Statistical Sciences. 1-3.

Ethical Considerations and Feedback from Social Human-Robot Interaction with Elderly People

Lucile Béchade*, **Agnes Delaborde***, **Guillaume Dubuisson Duplessis***, **Laurence Devillers***,**

*LIMSI–CNRS, Université Paris-Saclay, Orsay, France

**Université Paris–Sorbonne, Paris IV, France

{bechade, agdelabo, gdubuisson, devil}@limsi.fr

Abstract

Field studies in Human-Robot Interaction are essential for the design of socially acceptable robots. This paper describes a data collection carried out with twelve elderly people interacting with the Nao robot via a Wizard-of-Oz system. This interaction involves two scenarios implementable in a social robot as an affective companion in everyday life. One of the scenarios involves humour strategies while the other one involves negotiation strategies. In this paper, the authors detail the designs of the system, the scenarios and the data collection. The authors take a closer look at the opinions of the elderly collected through self-reports and through a private verbal exchange with one experimenter. These opinions include recommendations about the features of the robot (such as speech rate, volume and voice) as well as points of view about the ethical usage of affective robots.

Keywords: Social robotics, Elderly people, User feedback, Affective computing

1. Introduction

In assistive and social robotics, experimentation with potential end-users provide a valuable feedback about their expectations: through questionnaires and discussions, they give the researcher tracks to follow so as to improve the acceptability of the interaction system. These studies on the usage of robots are decisive for the co-conception of social Human-Robot Interaction (HRI) systems, which, by involving the participants in the designing process, can increase the level of trust towards companion and assistive robots.

The authors present in this study an experiment that took place in a Parisian living lab that receives elderly people for workshops around new technologies. The study carried out by the authors consists in assessing the acceptability of the interactions. The participants interacted with the robot Nao[1], in French, on scenarios about humor and negotiation.

In a first section, the authors offer an overview of existing ecological experiments. Next, they detail the settings of the experiment, the type of data collected, and the systems and scenarios. The third section presents the results of the questionnaires and discussions with the participants.

2. State of the Art: Ecological Data Collections

Ecological experiments take into account the relations between the individual and their environment. Forlizzi et al. (Forlizzi et al., 2004) highlight the importance of carrying ecological experiments so as to stimulate the emergence of natural behaviors in the participants, which allows carrying more thorough studies on the participant's perception and feelings. Experimental methodology in anthropology sets the framework for ecological studies, which allows the creation and development of products that integrate more intuitively in the dailylife of users (Bell, 2002).

At the present time, many participants in Human-Robot Interaction studies are not familiar with robotic devices, which can make it hard for them to react naturally to a situation that is radically new to them. So as to make sure that the participant feels at ease in the context, the best the experimenter can do consists in setting a realistic environment (generally, a domestic environment), and/or in relying on the participant's strong interest and curiosity in new technologies (or robotics in particular).

Overall, the scientific community tends to leave the protected and controlled zone of laboratories, so as to carry out field studies and get in contact with real potential users of their systems. Living Labs offer an interesting compromise between field studies and in-vitro laboratory studies: field studies present the potential flaw of making the experimental protocols hard to control and implement, while Living Labs offer facilities to re-create an ecological context while still being able to control the environment and keep all the necessary material at hand, thus providing the comfort of a laboratory study. Among interesting works on HRI experiments carried out in Living Labs, one can cite (Sasa and Auberge, 2014), an HRI data collection for the study of socio-affective gestural and speech markers; a one-month study on the acceptance of robots carried out by (Wu et al., 2014), where the participants came once a week for four weeks to interact with the robot; cloud computing for mobile robots in smart environments (Bonaccorsi et al., 2015); a study on the robot acting like a social intermediate between a dependent user and the smart home (Johnson et al., 2014); the design of a robot for fall prevention and protection for elderly people (Fischinger et al., 2016).

3. Description of the Experiment

3.1. Context

The experiment took place in the LUSAGE Living Lab (Pino et al., 2014) at the Broca hospital, Paris, under the supervision of the gerontology service. The Living Lab features regular workshops in the framework of the "Café Multimédia" project. This project aims to bridge the digital divide (the access to digital technology is not homogeneously distributed among the population) that can be observed in

[1]Humanoid robot Nao by Aldebaran Robotics, Paris.

elderly people, which could potentially increase their social isolation. They offer the participants to discover the Information and Communication Technologies, to discuss them, and to meet designers and researchers. This settles a win-win situation for both the participants and the researchers: the latter can put their ideas and systems to the critical analysis of the former, who are particularly eager to learn and to add their building blocks for the development of modern technologies.

The authors of this study took part in the workshop on social robotics for health-care and everyday life. On this occasion, they offered the participants to interact with the robot Nao, and discussed (individually and in group) with them, about the experiment, but also about their opinions on social and assistive robotics in general.

The preliminary phase of the experiment consisted in a general and collective explanation of the experiment: the aims of the researchers, the type of interaction and the nature of the robot they were going to meet. The volunteers signed and kept a copy of a written authorization that recapitulates the experiment, the identity of the research team, the way the data will be subsequently anonymized and used for research only, and a reminder of their right of withdrawal at any time.

The participants interacted individually with Nao in a separate room (see Figure 1). To foster a climate of trust between the participant and the researchers, an experimenter took the time to present each capture device used in the experiment: "Here is the microphone, we're going to record your voice with that: it's connected to this computer. On this other computer screen you can see the whole scene being filmed from this side, and this camera is for a close-up of your bust".

Twelve participants took part in the experiment, in total 8 males and 4 females, for a median age of 78. These participants did not present any relevant cognitive or physical disability which could alter their interaction with the robot. The participants were eager to discover and discuss about social robotics, but not familiar with such a technology. Indeed, only two participants out of twelve reported having previously interacted with a robot. These interactions were different from our experiment, insofar as they neither involved a social robot, nor did they concern a *verbal* interaction with a robot (i.e. which could communicate "in a human way"). In addition, the robots were presented to a group of people, not in a one-to-one social interaction.

3.2. Experimental Settings

The participant is asked to interact naturally on two scenarios: in a first scenario, Nao tries to make the participant laugh, by telling jokes or asking riddles; in a second scenario, Nao tries to negotiate for the participant to go and get him(her)self a glass of water. These two scenarios represent research themes of the team, consisting in studying how robotic humor is perceived, and the acceptability of a robot that suggests the user to do something trivial. The order of the scenarios was distributed equally among the participants.

The participant is filmed via a hand camera for a global recording of the scene, and a webcam for subsequent studies on facial emotion recognition. The voice of the participant is recorded through a headset directional microphone, for the team's research on emotion detection in speech. About one hour and a half of interaction has been recorded.

Figure 1: A participant introduces herself to Nao.

3.3. Scenarios

3.3.1. Humor Scenario

The humor scenario implements a system-directed entertaining interaction dialog that includes the telling of riddles and other humorous contributions. This scenario is meant for studying the impact and acceptability of the humor in robots.

All the interactions with the robot follow a common structure. First, the robot greets the participant and presents itself in an introduction phase. Next, the system offers the telling of riddles and jokes depending on the emotional state of the participant, which is perceptively assessed by the operator of the system. Then the robot challenges participants in a game by asking a question on cooking (e.g., "What ingredients are needed to cook an onion soup?"). Finally, the robot gives a conclusion on the participant's reactions: for example, if the participant seemed to have enjoyed the jokes, Nao will say "I am glad you like humor produced by a robot". Then the robot closes the interaction.

The humorous contributions of the robot are of the hackneyed variety (Bechade et al., 2016):

- Humorous riddle : e.g. "How do you know there are two elephants in your fridge? – You can't close the door."
- One-line joke : e.g. "This remind me of an anecdote: to fall asleep, a sheep can only count on itself."
- Teasing the participant: e.g. "Even a child could answer that!"
- Play on word : e.g. "Really, it's a piece of cake !"

3.3.2. Negotiation Scenario

In this setting, the robot is pictured as an assistive companion which will remind the user to drink some water.

This scenario aims at studying the social acceptability of the strategies used by the robot.

In a first step, the robot explains to the participant that it is willing to show him(her) the way it can negotiate for something. It gives one single instruction to the participant, which is that he(she) has to constantly refuse its offer.

When the scenario begins, the robot reminds the participant that it is time to re-hydrate, and that they would better get a glass of water. The robot does not offer to fetch itself the glass, nor does it make any movement in that sense. During this initialization phase, which lasts for two turns, the robot observes the way the participant reacts. This observation leads to the update of the user's profile (which will be detailed in the section below). Next, the robot selects a negotiation strategy according to the participant's detected profile, among:

- "humor": the robot makes derisive comments about itself so as to bring the user to accept its offer ("I can drink it for you. But I'm going to rust and you'll see that my motors can be far noisier than that get *really* noisy.")
- "appeal to reason": the robot presents some reasonable arguments ("Please, think about your health, it's for your own good.")
- "calming": the robot clearly establishes that it does not want to force the participant ("Calm down please, I don't mean to control you.")

The robot keeps the selected strategy during the last three turns of the scenario, with different sentences. In the end of the scenario, the robot reminds that it was only a simulation, and thanks the participant.

3.4. Systems

Two experimenters, present in the room, drive the two systems that trigger the predetermined utterances and gestures of the robot for each scenario. This present section details the systems: the Humor system is a Wizard-of-Oz (fully operated by a human experimenter), and the Negotiation system is semi-autonomous (the experimenter provide some of the inputs).

3.4.1. Humor

The humor scenario is driven by a Wizard of Oz system. The software designed for the team's experiments includes a graphic user interface allowing the human operator to remotely control the Nao robot.

It is configured by a predefined dialog tree that specifies the text utterances, gestures and laughter that can be selected to be executed by Nao. At each node, the operator chooses the next node of dialog to visit according to the human participant's reaction (Devillers et al., 2015).

The behavior of the system depends on the receptiveness of the human to the humorous contributions of the robot. Positive reactions (e.g. laughter, positive comments or positive emotions) lead to more humorous contributions, whereas repeated negative reactions (e.g. sarcastic laughter, negative comments and negative emotions) drive the dialog to a rapid end.

3.4.2. Negotiation

The experimenter enters manually the emotional expressions of the user (emotion type, intensity of the expressed emotion and speech duration). From these manual inputs, the emotional profile of the user is automatically updated, and drives the selection of negotiation strategy.

The profile is inspired by the Five-Factor personality model, introduced notably by (Hofstee et al., 1992; McCrae and John, 1992) and subsequently widely used for Human-System Interaction and Natural Language Processing. The profile deals specifically with the emotional components of the personality model (Delaborde and Devillers, 2010). The profile is composed of the emotional extroversion of the user (their propensity to react by expressing strong emotions), the emotional variability (an estimation of the variations between positive and negative expressed emotions), optimism (the ability to react positively) and self-confidence (an ability to cope).

Each strategy of the robot is associated to a specific lexical production and selected from the interpersonal circumplex (Delaborde and Devillers, 2012). The interpersonal circumplex, a theory based on interactional psychology concepts initially developed by (Leary, 1958), defines the social attitude of the interactant, on friendliness and domination axes ("with"–"against"; "above"–"below") .

The strategy of the robot is selected according to the user profile, based notably on the relationship between personality and interpersonal attitude (as defined, among many others afterwards, by Leary), and the notion of interpersonal complementarity between the two interactants (Carson, 1969). In the negotiation scenario presented in this study, the extroversion, self-confidence and optimism were parameters for the selection. For example, the robot would appeal to the participant's reason if the latter is globally self-confident, extrovert and optimist.

3.5. Adaptation of the Wizard-of-Oz Traditional Settings

The authors wish to note that in the settings of this data collection, the experimenters did not try to hide the fact that the robot was not entirely autonomous. Resorting to Wizard-of-Oz settings (i.e. making the participant fully believe the system is not piloted by a human) may prove useful for interactional data collection involving naive participants, so as not to cause a bias in their interaction with the system: the notion that the experimenter is making the decisions during the interaction could lead them to talk to the experimenter, rather than to the system. This notion is explained in (Dahlbäck et al., 1993), in which is detailed the interest of resorting to Wizard-of-Oz systems for dialogic interactions.

Nonetheless, in the context of this present encounter, the participants are involved in a collaborative reflection around the use of new technologies in daily life, and are deeply interested by the technical considerations around these fields. The authors felt that it would be more rewarding, both for the participants and for themselves, to give them the opportunity to dialog about the technical aspects of the systems, to involve them in the research process.

With these considerations in mind, one can understand that the experimenters found a comprise: on one side, they did not explicitly announce, at the outset, to the participants that the latter would not talk to an autonomous robot (the authors wanted to give them the opportunity to engage in interaction). On the other side, they made no particular effort at hiding their equipment (cables, computers), on the basis that this data collection was not only meant to collect interactional data, but also to bring about an exchange and confront the authors' research work with involved potential end-users.

Only one participant asked for confirmation that the experimenters controlled the robot, but the impact on the participant's naturalness is hard to assess in the present experiment.

4. Self-report Questionnaire and Verbal Exchange

4.1. Questionnaire

The written questionnaire, presented at the end of the experiment in another room, concerns the participant's overall acceptation of the interaction with the robot. The participants had to give their opinion on:

- Their general feeling toward the robot:
  - Did you feel like talking to the robot?
  - Did you understand what the robot said?
- The scenario Humor:
  - Did you enjoy the interaction?
  - Was the robot funny?
  - Did you feel... *(the participant assesses all the adjectives)* amused, hurt, surprised, stressed, enthusiastic, cowed, self-confident, embarrassed, involved, introvert, extrovert?
- The scenario Negotiation:
  - Did you enjoy the interaction?
  - Was the negotiation appropriate?
  - Did you feel... *(the participant assesses all the adjectives)* enthusiastic, self-confident, extrovert, even-tempered?
- The robot's attitude:
  - Was the robot more friendly or hostile?
  - Was the robot more reassuring or threatening?
  - Was the robot more humble or dominant?
- General opinion on robotics:
  - Have you already interacted with a robot? In what circumstances?
  - Do you feel like owning a robotic companion? Would you give it a special name?
  - Is there any comment you wish to add?

4.2. Verbal Exchange

After interacting with the robot and filling in the questionnaire, the participants were invited to talk about their experience through a private verbal exchange with one experimenter. This interview was completely informal, and it was left to the participant him(her)self to decide the topic he(she) wanted to address, without any restriction.

For example, five out of twelve participants have reported non-critical audibility problems, resulting in a difficulty to understand certain words pronounced by the robot. This has happened despite precautions that the authors took, namely: a quiet and isolated room, an increase in the volume of the voice of the robot and a decrease of the speech rate of the robot. Participants pointed out specific factors making the robot voice not so clearly audible, namely the fact that: (i) the voice was too nasal, and (ii) the voice was too monotone. They also insisted on the fact that the robot should have a non human-like voice.

The participants notably commented on their responses in the questionnaire, and broached their concerns or interests about robotics, which the experimenter noted down. This step of the experiment is consistent with the desire of the research team to take the time to listen and take into account the opinions and feelings of the participants (and potential end-users).

4.3. Results

The size of our sample does not allow statistically significant results on the models tested during the scenarios, but offers an interesting and detailed feedback about the usage of the robot by elderly people. Subjects were really involved and analytic, and made constructive comments.

4.3.1. Interaction appraisal

The participants assessed the interaction on Likert scales [1-5], where 1 means "I strongly disagree", and 5 "I strongly agree". They noted that they were globally really willing to interact with the robot (median 4, mode 5), that they understood it well (median 4, mode 5) and that they enjoyed the interaction (median 5, mode 5). On the whole, they considered the Humor scenario as very entertaining and the negotiation strategies appropriate, as shown by the detailed results presented in figure 2. One can note that the participant who found the robot neither funny nor negotiating appropriately at all (ID 3) did not enjoy the overall interaction, and expressed the fact that they expected to be

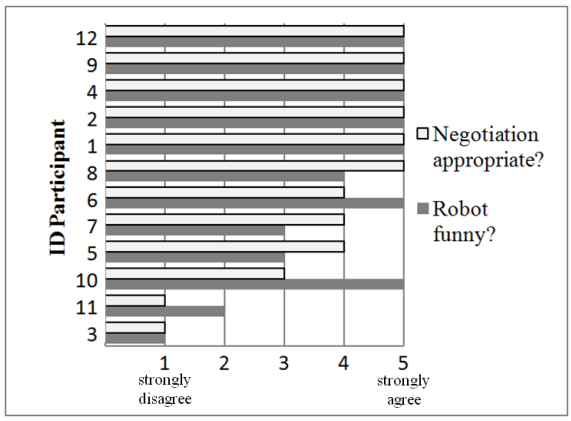

Figure 2: Answers from each participant to the questions "Was the robot funny?" and "Was the negotiation appropriate?".

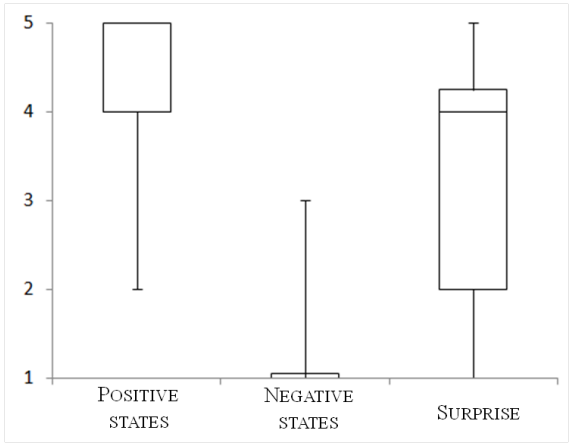

Figure 3: Reported positive, negative and surprise affective states of the participants (1 = "Not at all", 5 = "Very").

presented real assistive features.

The participants also reported on the socio-affective states they felt during the experiment. The Figure 3 presents the results of their self-report. One can note for example, that on the whole, participants experienced mostly positive states (among which enthusiasm, involvement). The negative states they could experience consisted in stress and embarrassment: the presence of the experimenters, the fact that they were being recorded, the novelty of the situation can potentially explain these feelings.

The Figure 4 presents the way the participants perceived the robot. For instance, on the Friendly–Hostile axis, they were expected to choose between: "Very friendly", "Quite friendly", "Not one more than the other", "Quite hostile", and "Very hostile". The same configuration applied for the axes Reassuring–Threatening and Humble–Dominant. While the robot was perceived, on the whole, as really friendly and reassuring, the participants highlighted the fact that a robot cannot be considered as "dominant". In their opinion, there is no notion of domination about a robot; the results mainly express their reservations about the consistency of such a term for a social robot.

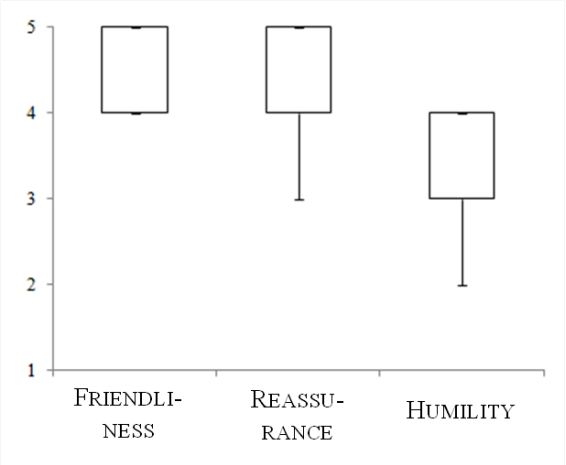

Figure 4: Reported perceived social attitude of the robot (1 = "Not at all", 5 = "Very").

4.3.2. Usage of Social Robots in everyday life

Participants have expressed significant opinions about the usage of social robots in everyday life.

After the interaction, participants were asked to report in the questionnaire whether they wanted to own a social robot or not. Mixed results were observed. Six participants reported being willing to own a social robot, four reported they were not, and two reported that they did not have a clear opinion at the moment.

The authors have also observed opposite opinions between participants about whether or not a robot should be involved in an affective relation with a human. Three out of twelve participants expressed a strong opinion about the fact that they do not want a robot to replace a human presence. These participants clearly stated their concern about the replacement of humans by a robot in affective relations. On the contrary, one participant highlighted the benefit of having a robotic companion in order to diminish the loneliness among isolated people. The other eight participants did not express a clear view about that subject.

The six participants willing to own a social robot did not express concerns about the long-term implication of robots replacing humans. On the other hand, three out of the four participants who are not willing to own a robot did express concerns about that matter. Besides, it seems that concerns about the involvement of a robot in an affective relation with a human is not caused by an unsatisfactory interaction in the course of this experiment. Indeed, the three concerned participants reported to be willing to interact with the robot and that they enjoyed the interaction.

Participants outlined specific use cases that they considered relevant for a robot. For example, they noted the benefit of having a robot in a context of personal assistance for people with disabilities or with disabling diseases in their daily life. While some participants have concerns about relations between a robot and a single human, they mentioned as interesting the usage of robot in interaction with a group of humans, such as a robot waiter in a family context.

The authors notice a discrepancy between the way the participants reacted in the course of the interaction, and the doubts and concerns they expressed during the verbal exchange. Indeed, the participants reported in the questionnaire that they felt satisfied with the robot's attitudes and pleased with the overall interaction, but their clearly decided opinions about social robotics transpired through the verbal exchange. The authors think that the experimental settings (i.e. they do not own the robot, they participate out of scientific curiosity) lead them to express affects which were partly due to the novelty of talking with a robot, and not only due to the nature of the interaction in itself.

5. Discussion and Conclusion

Studying natural interactions between humans and robots is primordial in the design of interfaces. However, the eagerness and surprise of participants towards robotic devices should temper the conclusions and tendencies computed from data and questionnaires. Though informal, verbal exchanges with participants allow the researcher to address many questions which could lead to an improvement of the robot's acceptability.

Some participants expressed their concerns about living with a robot that listens to everything; others felt put at an advantage that the society is interested in designing such technologies for them, and that they ask them about their opinion. As one of the participants declared: "I feel proud when I talk to my grand-children about these experiences". Group discussions allow the researchers to open up the debate on the designed and tested technologies, but also to explain more specifically their scientific challenges, progress and limitations.

The place of robots in the society gives rise to many concerns, such as the protection of data collected by the robots in a domestic context, the physical safety, the distribution of work, steering dependent people towards even more isolation, etc. The step of information, communication and popularization of science is crucial in experimental methodology in Human-Robot Interaction, and could contribute in an increased acceptation of the robots.

6. Acknowledgment

The authors wish to thank Pr. Anne-Sophie Rigaud, head of the gerontology department at the Broca hospital, and her team, for providing access to the LUSAGE Living Lab facilities, hence allowing the authors to carry out this study. The authors would also like to show their gratitude to all the participants of the experiment.

7. Bibliographical References

Bechade, L., Dubuisson Duplessis, G., and Devillers, L. (2016). Empirical study of humor support in social human-robot interaction. In *18th International Conference on Human-Computer Interaction (in press)*.

Bell, G. (2002). Looking across the atlantic: Using ethnographic methods to make sense of europe. *Intel Technology Journal, Q 3 ,*, 1–10.

Bonaccorsi, M., Fiorini, L., Cavallo, F., Esposito, R., and Dario, P. (2015). Design of cloud robotic services for senior citizens to improve independent living and personal health management. In *Ambient Assisted Living*, pages 465–475. Springer.

Carson, R. C. (1969). *Interaction concepts of personality.* Aldine Publishing Co.

Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). Wizard of oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*, pages 193–200. ACM.

Delaborde, A. and Devillers, L. (2010). Use of nonverbal speech cues in social interaction between human and robot: emotional and interactional markers. In *Proceedings of the 3rd international workshop on Affective interaction in natural environments*, pages 75–80. ACM.

Delaborde, A. and Devillers, L. (2012). Impact of the social behaviours of the robot on the user's emotions: Importance of the task and the subject's age. In *WACAI 2012 Workshop Affect, Compagnon Artificiel, Interaction*, page 167.

Devillers, L., Rosset, S., Duplessis, G. D., Sehili, M. A., Béchade, L., Delaborde, A., Gossart, C., Letard, V., Yang, F., Yemez, Y., Türker, B. B., Sezgin, M., Haddad, K. E., Dupont, S., Luzzati, D., Esteve, Y., Gilmartin, E., and Campbell, N. (2015). Multimodal data collection of human-robot humorous interactions in the joker project. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 348–354, Sept.

Fischinger, D., Einramhof, P., Papoutsakis, K., Wohlkinger, W., Mayer, P., Panek, P., Hofmann, S., Koertner, T., Weiss, A., Argyros, A., et al. (2016). Hobbit, a care robot supporting independent living at home: First prototype and lessons learned. *Robotics and Autonomous Systems*, 75:60–78.

Forlizzi, J., DiSalvo, C., and Gemperle, F. (2004). Assistive robotics and an ecology of elders living independently in their homes. *Journal of HCI Special Issue on Human-Robot Interaction.*, 19:25–59.

Hofstee, W. K., De Raad, B., and Goldberg, L. R. (1992). Integration of the big five and circumplex approaches to trait structure. *Journal of personality and social psychology*, 63(1):146.

Johnson, D. O., Cuijpers, R. H., Juola, J. F., Torta, E., Simonov, M., Frisiello, A., Bazzani, M., Yan, W., Weber, C., Wermter, S., et al. (2014). Socially assistive robots: a comprehensive approach to extending independent living. *International journal of social robotics*, 6(2):195–211.

Leary, T. (1958). Interpersonal diagnosis of personality. *American Journal of Physical Medicine & Rehabilitation*, 37(6):331.

McCrae, R. R. and John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

Pino, M., Benveniste, S., Picard, R., and Rigaud, A.-S. (2014). User-driven innovation for dementia care in france: The lusage living lab case study. *Interdisci-*

plinary Studies Journal, 3(4):251.

Sasa, Y. and Auberge, V. (2014). Socio-affective interactions between a companion robot and elderly in a smart home context: prosody as the main vector of the" socio-affective glue". In *SpeechProsody 2014*.

Wu, Y.-h., Wrobel, J., Cornuet, M., Kerhervé, H., Damnée, S., and Rigaud, A.-S. (2014). Acceptance of an assistive robot in older adults: a mixed-method study of human-robot interaction over a 1-month period in the living lab setting. *Clinical interventions in aging*, 9.

Translation Resources and Translator Disempowerment

Joss Moorkens[1], David Lewis[2], Wessel Reijers[1], Eva Vanmassenhove[1], Andy Way[1]

[1]ADAPT Centre, School of Computing, Dublin City University, Ireland,

[2]ADAPT Centre, Trinity College Dublin, Ireland

E-mail: joss.moorkens@dcu.ie, dave.lewis@adaptcentre.ie, wreijers@adaptcentre.ie,

eva.vanmassenhove2@mail.dcu.ie, away@computing.dcu.ie

Abstract

Language resources used for machine translation are created by human translators. These translators have legal rights with regard to copyright ownership of translated texts and databases of parallel bilingual texts, but may not be in a position to assert these rights due to employment practices widespread in the translation industry. This paper examines these employment practices in detail, and looks at the legal situation for ownership of translation resources. It also considers the situation from the standpoint of current owners of resources.

Keywords: Language resources; copyright; ethics

1. Introduction

Statistical Machine Translation (SMT: Koehn et al., 2003, 2007), the most prevalent paradigm currently for automatic translation, requires large amounts of bilingual parallel language resources. These resources are originally created by human translators whose rights with regard to their creation are not always respected, and who are disempowered by the vendor model widespread within the language services industry. While the proportion of freelance or contingent workers in developed countries has increased, reaching 40.4% in the US in 2010 (U.S. Government Accountability Office, 2015), surveys of translators have found the proportion of freelance workers to be in the region of 80% (84% in Kelly DePalma, & Hegde, 2012; 77% in Ehrensberger-Dow et al., 2015). This prevalence of freelance translation has the effect of disempowering translators who otherwise might be in a position to assert their copyright for work created or derived work, as well as for collective bargaining for pay rates and conditions.

The issue of falling pay rates and of unclear ownership of translation databases has grown in prominence as the use of Translation Memories (TMs: Heyn, 1998) as repositories for previously translated work has become widespread, in particular as 'fuzzy match' scoring (Sikes, 2007) against client-provided TMs has become a common discounting mechanism in pricing translation projects. TMs are also the primary source of parallel text for the training of SMT engines. The use of SMT is already widespread in translation projects, so the opportunity for effective leverage of data from a specific TM in translating content from different clients or domains has widened.

In this paper we look in more detail at translators' employment conditions and their association with practices prevalent in the language industry with regard to data ownership. We then examine how copyright might apply to translated texts and TM databases, and finally offer some recommendations from the perspective of translators and for regulation of copyright ownership.

2. Translators' Agency

Translators' have found their profession increasingly limited in several ways: conditions of employment have moved to a freelance model with an associated loss of security and benefits, technologization of the translation industry has reduced translator autonomy, and the related move to the digital domain has made the situation unclear with regard to the ownership and reuse of translated material. In the following sections, we examine each of these issues in turn.

2.1 The Vendor Model

From a high point of "de-commodification" of labour and gains in worker power in the boom years post-WW2 (Munck et al., 2011), many industries have moved to a freelance model, where workers have become self-employed contractors, who have to "buy their own tools and equipment, and bear all the risks of accident, sickness, or lack of work" (Castles, 2011). The translation industry has, to a great extent, moved in this direction. Reliance on freelance translation work has become widespread among language-service providers, as the freelance model is "flexible, scalable, or cost-effective enough to respond to market demands" (Kelly et al., 2012). This may allow translators a degree of autonomy, but for most translators outside of those working for larger public institutions, there is little choice, especially if they wish to continue to translate rather than moving into management within a company. A survey by Moorkens and O'Brien (2016) found an association between translators' age ranges and their working conditions, where those over the age of 30 are far more likely to work on a freelance basis. Many freelancers (31% of the total) work directly for one agency, a situation referred to as *bogus self-employment* in a study of precarious work for the European Commission. When a freelancer's relationship is "with a single source rather than with a range of clients", this represents "economically dependent work" (McKay et al., 2012).

This situation leaves translators in a difficult position with regard to collective bargaining, negotiation of rates, and

assertion of copyright. Even though the language industry has continued to show year-on-year growth of over 5% through the recent recession (DePalma et al., 2013), freelance translators have complained of their powerlessness in the face of shrinking per-word rates that are often dictated by their agencies (Kelly, DePalma, & Hegde, 2012).

2.2 Translators and Technology

Translator disempowerment has been exacerbated by the technologization of the translation profession since the introduction of TM technology in the early 1990s. While some translators have been early adopters of new technologies, many resent that new technologies are imposed on them (Penkale & Way, 2013; Way, 2013): first TM with its associated fuzzy match discounts, and more recently MT post-editing, which requires them to accept further discounted rates to fix "fundamental linguistic errors that a trained human translator would rarely generate" (O'Brien, 2012). It is rarely made explicit by companies and research groups that specialize in MT that human translation is its necessary basis, with the focus instead on new and better ways to process this trove of pre-existing 'big data' (Kenny, 2011). The gradual limitation of the translator's role has undermined their ability to conform to the ethical code of their profession (Chesterman, 2001) by reducing the translation process to a series of "language-replacement exercises" (Pym, 2003). Furthermore, as the profession has moved from analogue to digital, translators' powerlessness is reflected in continued data dispossession, common for many knowledge workers, and largely unaffected by legal constraints (Huws, 2014). This is a wider problem within the digital domain, where national laws are of little relevance, and assignation of rights is often buried within data-use policies (Reijers et al., 2016).

3. Ownership of Language Resources

Translators typically create a TM file as a by-product of a translation effort. Currently, handing over TM files to an agency after a translation job has become the norm in the translation industry, whether or not ownership has been specified in translation project contracts. In the absence of a contractual agreement regarding ownership of what Smith (2008) has called the "translation family jewels", the actual legal status of a translation or translation artefacts is subject to a variety of often ill-defined national and international laws and is thus unclear (Lewis et al., 2016). For example, authorship of a source text, including the right to decide whether work is translated, may belong to either the employer or employee depending on the country in which the author is contracted, and contractual assignment of authorship is only valid in some jurisdictions (Troussel & Debussche, 2014).

Unless specified in a contract, a translator may be considered the owner of a translated text as a derivative or adapted work, depending on the perceived originality of the translation and subject to the "rights of the author of the original work" (Troussel & Debussche, 2014). In the US, the claimant of copyright must demonstrate a "minimum degree of creativity" (Cabanellas, 2014). This situation becomes more complex when applied to user-generated content or crowd-sourced translation, for which no specific legal framework exists. The copyright for a database, such as a TM file, is considered to belong to the database creator in both France and Germany, depending on the originality involved with its creation, in this case regarding "segmenting and aligning the data" (Troussel & Debussche, 2014). There may be the option of asserting further *sui generis* rights to the creation of a database, if the creator has demonstrated a substantial investment in obtaining, verifying, or presenting that database (Troussel & Debussche, 2014).

The situation with regard to copyright issues internationally appears fluid. Copyright laws have changed over time in many jurisdictions, and within the EU are further complicated by a number of EU-level directives that are intended as a step to harmonize copyright , and to address new issues raised by unexpected technological advances, permitting mass digitization of books, for example. Periodic public consultations have taken place, most recently in 2013, which look to address issues with text and data mining, and user-generated content, and have been followed up with the establishment of European Commission working groups.[1]

The somewhat fluid state of copyright law has not appeared to effect the reality for ownership of translation data, which (to our knowledge) has never been legally tested. Freelance translators continue to deliver TMs to their client or agency without question, as the failure to do so may affect the "translator's standing with that service provider" and "payment problems could ensue" (Smith, 2008). This situation is critical especially for the large proportion of translators who work directly with a single agency.

3.1 Consequences for Reuse

Although these potentially conflicting claims of copyright for written or translated material are currently ignored, they may create difficulties for enterprises offering MT and, to a lesser extent, collectives sharing MT. For translators, the re-tasking of TM as parallel text for training MT engines is a particular concern (Moorkens & O'Brien, 2016).

The leverage of TMs from previous translations is well understood by translators. They understand the role it plays in avoiding unnecessarily retranslation of similar segments and the resulting role played by matching scores between available TMs and the source of incoming translation projects in price discounting. The practice of individual translators retaining TMs from previous projects independently of vendors is widespread, as modern desktop

[1] See Text and Data Mining Working Group website at: https://ec.europa.eu/licences-for-europe-dialogue/en/content/text-and-data-mining-working-group-wg4.

translation tools allow them to use these as reminders of previous translations and for term concordancing. These are useful features for individual translators even if the level of useful TM matching leverage with a personal TM is low. These practices seem to indicate a tacit approval by translators of the use of TM leverage. There seems to be an appreciation that they benefit from the prior work of other translators captured in a TM in the same way that other translators will benefit from their work in future. We can assume there is a degree of collegiality at play here, since even if translators producing and consuming translation via TM may not know each other's identities directly, the poor level of TM leverage across domains or client content types means benefitting translators can be assumed to be working in the same broad domain as those who produced the content.

The use of TMs for MT training erodes this traditional acceptance of TM leverage, since translators perceive that the resulting MT system can be used by vendors and clients for translation in very different domains. In particular MT is seen to be useful in classes of translation tasks where little or no translator input is required (cf. Way, 2013), contributing to the misconceived perception that the spread of MT endangers the livelihood of translators.

Although TM data interoperability standards, such as Translation Memory eXchange (TMX) [2] and XML Localization Interchange File Format (XLIFF) [3] enable translator provenance to be recorded, such metadata is typically stripped from TMs before being returned to clients or used between projects by vendors. The traditional acceptance of TM leverage means that, outside of a specific translation project, the tracking of the provenance of individual translation to specific translators is not practised, and is not strongly demanded by translators. However, the loss of this provenance data means that there is no way for individual translator contributions to large aggregated TMs to be differentiated, and hence translators are denied the opportunity to specify any preferences on the rights they wish to declare over the use of TMs they return to vendors and clients.

The situation in Public Service Institutions, with regard to the collection and sharing of resources, may be somewhat simpler, depending on where they were created. The EU has a harmonized directive for re-use of information that was enacted in 2003 (directive 2003/98/EC)[4] and updated in 2013, which stipulates that written texts, databases, audio files and film fragments held within public repositories (with some exceptions) may be reusable for commercial and non-commercial purposes. These purposes need not relate to the initial intended purpose of the data. The only difficulty remaining in this instance is whether the data was created by an external party, in which case they may not have been made aware of the Public Service Information directive, nor have supplied materials such as parallel data that were created during the process of completion of their task. In this case, there may be a requirement to negotiate the release of data ownership retrospectively.

The situation at present in which laws of copyright are effectively bypassed in content collection, curation, and exploitation, permits resource holders to retain data at a cost to disempowered human writers and translators, and also at a cost to end-users of translated content. The disconnect between the MT services and the human translated corpora might further alienate translators from their work, and add to existing mistrust in MT and in data sharing.

4. Recommendations

Working largely independently within the vendor model with increasing imposition of translation technology, there are nonetheless possibilities for freelance translators to maximize their agency through collective bargaining. This could be via a national or international translators' organization such as FIT (The International Federation of Translators)[5] or online groups such as proz.com.

The growing number of precarious workers in all industries – especially for well-publicized technology companies using a crowdsourcing model such as Uber and Amazon's Mechanical Turk – has made precarious work a topical issue. 30% of paid jobs in the EU between 1987 and 2007 were temporary work, and the percentage of flexible employment contracts issued in Greece rose from 21% in 2009 to 41% in 2011 (McKay et al., 2012). In the US, the number of contingent employees more than doubled between 1969 and 1993 (Cummings & Kreiss, 2008). Concern over this issue has led to sporadic moves to allow contingent workers the right to organize, with legislation for limited collective bargaining on behalf of freelance workers progressing towards being enacted in law in Ireland in 2016 (Houses of the Oireachtas, 2016), and collective bargaining agreements are already in place for several categories of contingent workers in Washington State since 2013. One of these categories of workers is Language Access Providers, defined as 'any independent contractor who provides spoken language interpreter services for Department of Social and Health Services appointments or Medicaid enrollee appointments' (Washington Federation of State Employees, 2013). This bargaining agreement defines rates of pay, payment deadlines, and a grievance procedure. If these agreements are considered successful, there may be grounds for expanding to other categories and professions.

A second recommendation for translators is to inform themselves about their legal rights for translation. This could be encouraged via a conversation in the wider language service industry, and volunteer translation organizations such as Translators Without Borders[6] and The Rosetta Foundation[7] could also raise awareness by

[2] https://www.gala-global.org/tmx-14b
[3] http://docs.oasis-open.org/xliff/xliff-core/v2.0/xliff-core-v2.0.html
[4] http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32003L0098
[5] http://www.fit-ift.org/
[6] http://translatorswithoutborders.org/
[7] http://www.therosettafoundation.org/

explicitly using an open or standard data ownership policy and allowing volunteers to control the ways in which the content that they translate is leveraged.

A third recommendation is that translators use TM metadata more effectively to both identify the translations and translation alignments in which they had a creative input and to explicitly assign usage rights to those assets. While such metadata can be captured in existing TM data standards (TMX and XLIFF), population and maintenance of this metadata needs to be integrated into translation workflows. In addition, better shared models for differentiating used of assets is required. For example, Lewis et al. (2016) suggest an extension to the existing metadata vocabulary for expressing usage rights to allow differentiated usage rights between traditional TM leverage and TM use in MT training to be declared. A clear and legally defensible definition to allow this differentiation to be unambiguously established in any given case is still required, however.

Recent efforts to harmonize copyright laws in the EU are welcome and any agreed ethical code for collection and reuse of human translations will need to be universally agreed. The potential financial implications of this in an industry valued at US$34.778 billion in 2013 (DePalma et al. 2013) are likely to make agreement difficult to achieve.

5. Conclusion

The prevalence of the vendor model in the translation industry shows no sign of abating. As noted by Linder (1999), once cost-cutting employment practices become commonplace in an industry, other players are pushed into following those same practices in order to remain competitive. This does not necessarily mean that the outlook for translators is poor. The industry continues to grow, and governments and society are beginning to realize that they need to legislate for the protection of contingent workers and to allow collective bargaining.

Steps towards harmonization of copyright laws are being made, but legislation is particularly uneven in the digital domain, where working groups and consultations are taking place in an effort to keep up with technological changes. These developments are likely to have significant ethical implications for people working in the translation industry.

For translators, it is in their best interests to act collectively where possible, to maximize bargaining power and to share information, particularly with regard to making best use of the metadata possibilities of current interchange formats. Ideally, any agreement for collection, ownership, and reuse of translation data will come about via consensus, but more empowered translators may become emboldened to pursue copyright claims as described in Section 2, as a legal challenge on behalf of a translator could have massive repercussions in an industry where the norm is usually unchallenged.

Acknowledgements

This work has been supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund, by the European Commission as part of the FALCON project (contract number 610879), and by the Dublin City University Faculty of Engineering & Computing under the Daniel O'Hare Research Scholarship scheme.

6. References

Cabanellas, G. (2014). *The Legal Environment of Translation*. Abingdon: Routledge.

Castles, S. (2011). Migration, Crisis, and the Global Labour Market, *Globalizations*, 8(3), pp. 311—324.

Chesterman, A. (2001). Proposal for a Hieronymic Oath, In Pym, A. (Ed.), *The Translator*, 7(2), pp. 139-154.

DePalma, D. A., Hegde, V., Pielmeier, H., and Stewart, R. G. (2013). *The Language Services Market: 2013 (Report)*. Common Sense Advisory, Boston MA.

Heyn, M. (1998). Translation Memories – Insights & Prospects. In L. Bowker, M. Cronin, D. Kenny and J. Pearson (Eds.) *Unity in Diversity? Current Trends in Translation Studies*, Manchester: St. Jerome, pp. 123—136.

Huws, U. (2014). *Labor in the Global Digital Economy: The Cybertariat Comes of Age*. New York: Monthly Review Press.

Kelly, N., DePalma, D. A., and Hegde, V. (2012). *Voices from the freelance translator community (Report)*. Common Sense Advisory, Boston MA.

Kenny, D. (2011). The ethics of machine translation. In *Proceedings of the New Zealand Society of Translators and Interpreters Annual Conference 2011*. Auckland, New Zealand.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, M., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. *ACL 2007: proceedings of demo and poster sessions,* Prague, Czech Republic, pp.177—180.

Koehn, P., Och, F.J., and Marcu, D. (2003). Statistical phrase-based translation. In *HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, Edmonton, Canada, pp. 48–54.

Lewis, D., Fatema, K., Maldonado, A., Walshe, B., and Calvo, A. (2016). Open Data Vocabularies for Assigning Usage Rights to Translation Memories. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Linder, M. (1999). Dependent and Independent Contractors in Recent U.S. Labor Law: An Ambiguous Dichotomy Rooted In Simulated Statutory Purposelessness.

Comparative Labor Law & Policy Journal 21, pp. 187—230.

McKay, S., Jefferys, S., Paraksevopoulou, A., Keles, J. (2012). *Study on Precarious work and social rights: Carried out for the European Commission*. Working Lives Research Institute, London Metropolitan University.

Moorkens, J., O'Brien, S. (2016). Assessing User Interface Needs of Post-Editors of Machine Translation. In D. Kenny (Ed.), *Human Issues in Translation Technology: The IATIS Yearbook.* Abingdon: Routledge.

Munck, R., Schierup, C. U., and Wise, R. D. (2011). Migration, Work, and Citizenship in the New World Order, *Globalizations,* 8(3), pp. 249--260.

O'Brien, S. (2012). Translation as human–computer interaction. *Translation Spaces*, 1, pp. 101–122.

Penkale, S., and Way, A. (2013). Tailor-made Quality-controlled Translation. In *Proceedings of Translating and the Computer* 35, London, UK, 7pp.

Pym. A. (2003). Translational Ethics and Electronic Technologies, In *Proceedings of the Profissionalização do Tradutor Conference*. Lisbon, Portugal.

Reijers, W., Vanmassenhove, E., Lewis, D., Moorkens, J. (2016). On the need for a global declaration of ethical principles for experimentation with personal data. In *Proceedings of the ETHI-CA 2016 Workshop,* Portorož, Slovenia.

Sikes, R. (2007). Fuzzy matching in theory and practice. *Multilingual*, 18(6):39 – 43.

Smith, R. (2008). Your Own Memory. *The Linguist,* 47(1), pp. 22—23.

Troussel, J. C., Debussche, J. (2014). *Translation and Intellectual Property Rights* (Report by Bird & Bird for the European Commission DG Translation). Luxembourg: European Commission. doi:10.2782/72107

Way, A. (2013). Traditional and Emerging Use-Cases for Machine Translation. In *Proceedings of Translating and the Computer* 35, London, UK, 12pp

Ethics for Crowdsourced Corpus Collection, Data Annotation and its Application in the Web-based Game iHEARu-PLAY

Simone Hantke[1,2], Anton Batliner[1,3], and Björn Schuller[1,4]
[1] Chair of Complex & Intelligent Systems, University of Passau, Germany
[2] Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[3] Pattern Recognition Lab, FAU Erlangen-Nürnberg, Germany
[4] Department of Computing, Imperial College London, UK
Email: simone.hantke@tum.de

Abstract

We address ethical considerations concerning iHEARu-PLAY, a web-based, crowdsourced, multiplayer game for large-scale, real-life corpus collection and multi-label, holistic data annotation for advanced paralinguistic tasks. While playing the game, users are recorded or perform labelling tasks, compete with other players, and are rewarded with scores and different prizes. Players will have fun playing the game and at the same time support science. With this modular, cross-platform crowdsourcing game, different ethical and privacy issues arise. A closer look is taken on ethics in recording of private content, data collection, data annotation, and storage, as well as sharing the data within iHEARu-PLAY. Further, we address the interplay of science and society in ethics and relate this with our application iHEARu-PLAY.

Keywords: Crowdsourcing, Corpus Collection, Data Annotation, Ethics, iHEARu-PLAY

1. Introduction

Crowdsourcing – the process of distributing tasks to an open, unspecified group of people via the internet – is an arising collaborative approach in the area of speech and language processing; it can be harnessed for many different types of applications and offers instantaneously access to populations with specific knowledge and skills, everywhere on the globe, and for any spoken language. For annotating speech data, many projects employed crowdsourcing to save costs compared to expert human annotation in labs (Burkhardt et al., 2010; Zhai et al., 2013). Most crowdsourcing services rely on so called 'click-workers', which are being paid a rather low compensation for their work. Their jobs are often not very appealing and thus, their intrinsic motivation will be low. As is the case for most newly developed techniques, crowdsourcing also raises both hopes and doubts, certainties and also many questions. Eskenazi et al. (2013) give a general analysis of crowdsourcing for speech processing.

When dealing with crowdsourcing, many economic and ethical problems arise, which are related to the type of crowdsourcing services, the task to be addressed, the country where the click-workers perform the tasks, and the pertinent labour laws. Ethics is often equated with decisions of high moral magnitude and associated with weighty concepts of right and wrong. Although the relevance of ethics to daily experience is not always easy to assess, Seedhouse (1998) proposes a definition, highlighting this daily relevance by referring to ethics as a process of deliberation about "how best to conduct one's life in the presence of other lives" – in fact, a sort of reformulation of Kant's categorical imperative: "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law.". In fact, this simple statement refers to a very complex multivariate problem, and solutions should be found to deal with crowdsourcing services in a more efficient and ethical way.

This paper presents an alternative way to collect annotated or recorded data by crowdsourcing non-professional subjects with the fun gaming platform iHEARu-PLAY (Hantke et al., 2015). iHEARu-PLAY motivates people by giving them a playful environment, where they can have fun and at the same time, voluntarily help scientific research projects by annotating data or recording prompted tasks. Instead of offering a financial incentive, people are primarily motivated to participate due to the joyful experiences of a game. Usually, the motivation for an individual to voluntarily contribute to a crowdsourcing project ranges from altruism, over ego, to a shared sense of purpose; yet, the pursuit of fun and enjoyment through games is also seen as an emerging trend (Good and Su, 2011). On top of the intrinsic motivation of playing a game and helping science, various prizes and awards can also be given, e. g., for the best scores, the most frequent users, and/or for randomly selected winners.

The present topic belongs to the broader field of Computational Paralinguistics (Schuller and Batliner, 2014; Batliner and Schuller, 2014). It seems that studies addressing ethics in connection with crowdsourcing so far dealt with paid work and not with the kind of voluntary work we are aiming at, cf. Silberman et al. (2010) and Schmidt (2013); Adda and Mariani (2013) address economic, legal and ethical aspects with crowdsourcing for speech, Adda et al. (2014) crowdsourcing in the context of big data.

Section 2 describes iHEARu-PLAY's concept and main idea. The ethical and personal issues are addressed in Section 3, including the types of data collected and how it will be used. In Section 4, we deal with general considerations, before summarising this paper in Section 5.

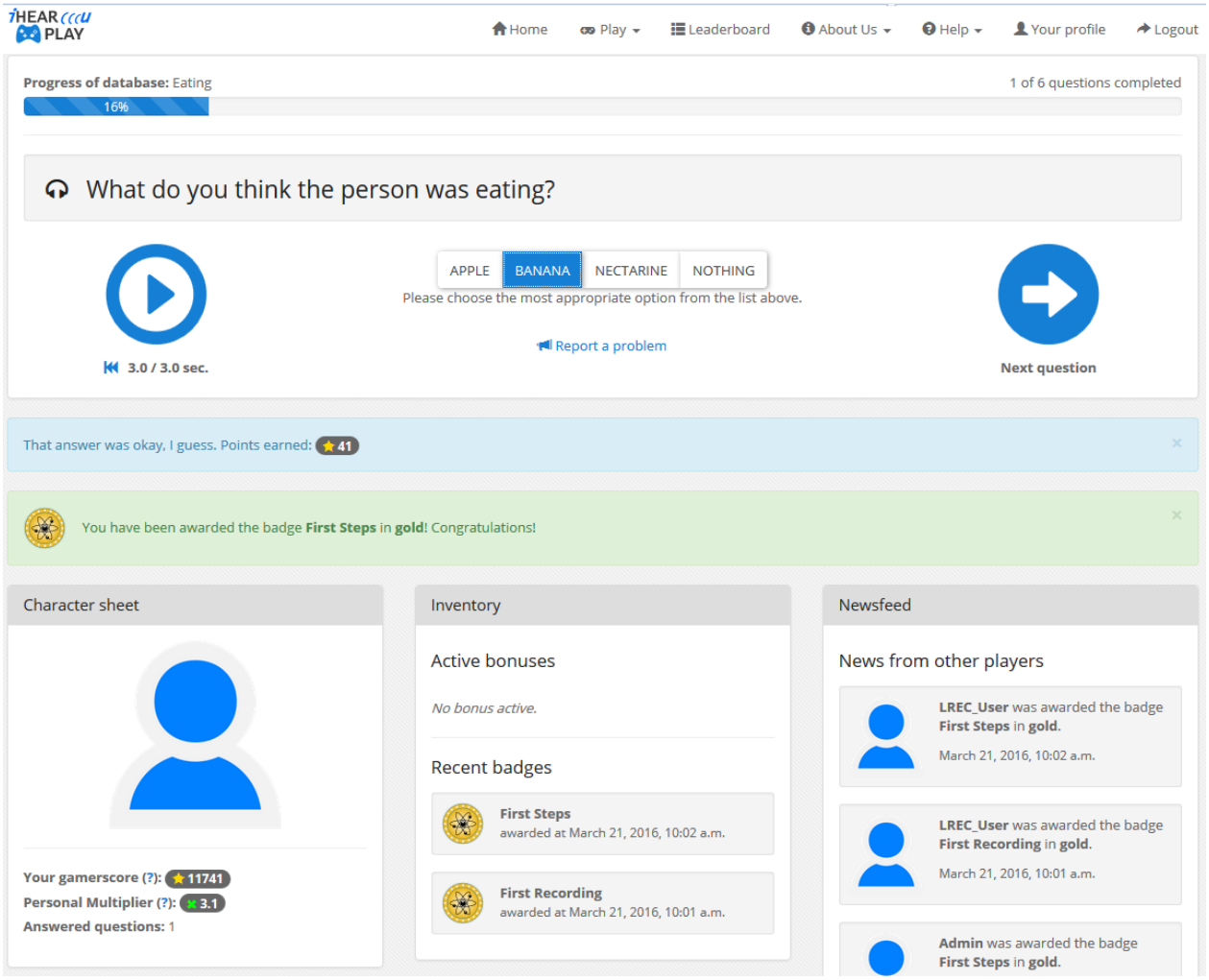

Figure 1: Exemplary screenshot: iHEARu-PLAY's web-interface for labelling tasks, showing the progress for the database, list of answers, the feedback message after submitting the answer, including the earned points, followed by the next audio file and question.

2. Our Work – iHEARu-PLAY

iHEARu-PLAY[1] is a web-based multiplayer game for crowdsourced database collection and data labelling (Hantke et al., 2015); its primary purpose is holistic, multi-label annotation of multi-modal affective speech databases usually containing also or only speech. Existing speech and video databases, and also image databases can be added, and labelling tasks can be defined via a web-interface. Further, new speech data can be collected by players performing prompted recording tasks in the wild. Players perform these labellings or prompted recording tasks and are rewarded with prizes and scores based on the 'correctness' of their annotations, e. g., the agreement with a pre-defined gold standard (an already existing annotation from a former lab annotation task) or the agreement with the (majority vote of the) other players.

When a new user visits iHEARu-PLAY for the first time, (s)he will be able to access a demonstration that explains the idea behind iHEARu-PLAY and teaches the user how to interact with the system. After signing up at iHEARu-PLAY, the player can choose a database for the annotation or recording task. Having picked an annotation database, the user will be presented with a random audio file and a question from that database. After playback of half of the audio file, a list of answers fades in from which the user can select one (or sometimes multiple). Having selected his or her answer, a submit button will fade in, which, upon execution, immediately presents a feedback message (based on the players performance), including the earned points as well as the next audio file and question. Then, the whole process starts over again. If the user earned a badge, it will be displayed in the same area as the feedback message and – if the user allowed to share his or her activity on the platform – on the activity ticker, thus visible to all other players. Figure 1 shows the web-interface for such a labelling task.

The web-interface for the recording task is build up similarly as the labelling web-interface and just differs from it in the small substituted part shown in Figure 2. Users can start the recording by clicking on the microphone, read the above shown sentence out loud, and get presented with a live spectrogram visualising their recorded speech.

After the user stopped the recording, (s)he can listen to the recorded prompt, if wished record the same sentence again and finally upload the recorded prompt to iHEARu-PLAY.

[1] https://ihearu-play.fim.uni-passau.de/

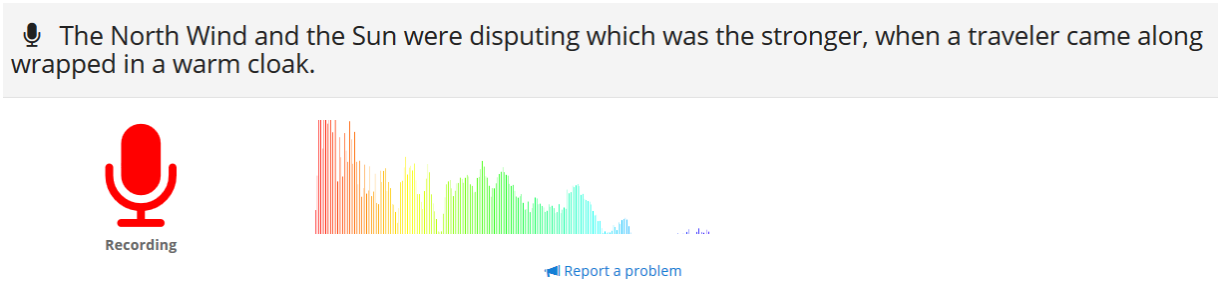

Figure 2: Exemplary screenshot: iHEARu-PLAY's web-interface for recording tasks, showing the microphone to start and stop the recording and the spectrogram to visualise the recorded speech of the user.

Figure 3: Exemplary screenshot: iHEARu-PLAY's web-interface for recording tasks after having recorded a sentence, showing the possibility to listen to the recorded audio file again, discard it or upload it to iHEARu-PLAY, followed by the next sentence.

Figure 3 shows the according part of the web-interface.
iHEARu-PLAY is implemented with the open source high-level PythonWeb framework Django (Foundation, 2014 Version 17) and can be installed on Unix and Windows platforms. Its modular architecture allows for easy integration of custom extensions: New gaming components can be added as plugins in order to support new databases and modalities. The game will be available to the research community as a ready-to-use web-service. Researchers can add their own databases, optionally post rewards, and receive annotation results in the end. General users can register to play the game, record new data or annotate already existing data, gain points for the tasks performed, compete with other players on the leaderboard, have fun, and at the same time support science (Hantke et al., 2015).

3. Privacy and Ethical Issues

An absolutely fundamental step is to determine the 'status' of users in a participant agreement. Many crowdsourcing platforms include in their terms of use a statement that defines the workers as 'independent contractors'. For instance, the Amazon Mechanical Turk Participation Agreement contains the statement: "As a Provider, you are performing Services for a Requester in your personal capacity as an independent contractor and not as an employee of the Requester." (Turk, 2016). In contrast to the typical crowdsourcing platforms, iHEARu-PLAY does not define the users as independent contractors but as volunteers. There is no monetary compensation for the tasks performed; iHEARu-PLAY is free of charge and only asks for voluntary participation.

For research scenarios, where data are collected, participating volunteers have to give informed consent. The iHEARu-PLAY informed consent form is included as an appendix to its platform. To ensure data anonymity and security, iHEARu-PLAY gives different restrictions and prohibitions within this form for users, e. g., an age restriction, the prohibition to give away personal information of themselves or other persons within the recordings or anywhere else on the platform, or generate unethical or inappropriate data (e. g., issues related to sexual or propaganda content). Within iHEARu-PLAY, there is a possibility for users to report other users if they generate or publish data against the privacy police or the general terms and conditions of iHEARu-PLAY. To avoid abuse, a fair, understandable and open concept of data collection, storage, usage, and sharing was developed for iHEARu-PLAY, which will be described in the following.

3.1. Collected Data

From a user perspective, 'privacy' is a highly nuanced, culturally pre-determined and context-dependent social concept. An activity that is entirely acceptable and appropriate in one context might not be acceptable in another context. Eventually, the user's own feelings and judgements have to be considered as guideline. Further, the idea that 'the internet never forgets' is extremely disturbing, given all the possible future uses of personal data. Therefore, it is ab-

solutely necessary to present an open and understandable concept of data collection to the user.

3.1.1. Data Types

Personally Identifiable and Mandatory Information: Many companies assure their customers or their users of the service that collected personal data will be released only in a non-personally identifiable form. The underlying assumption is that 'personally identifiable information' is a fixed set of attributes such as names and contact information. iHEARu-PLAY goes one step further and will not ask the user for a name or address in the first place. Nevertheless, in order to use all functions of iHEARu-PLAY, users must register to the platform first. For this purpose, providing basic data is necessary, such as a freely chosen username, an associated e–mail address, and a password. Further, iHEARu-PLAY saves a user's log-file for a duration of maximum seven days. This log-file will be deleted automatically after the given time and will just be used in cases of technical issues after being contacted by the user. Since the IP–Address of a user can easily be freely chosen and will not be stored for long time, the only traceable information extracted and stored could be the e–mail address, if the user's actual name is encoded there.

Anonymous and Voluntary Information: In addition, further data might be disclosed voluntarily in the personal profile of iHEARu-PLAY, e. g., a user's age, gender, personal health issues etc.; this is marked as optional information which – under certain circumstances – also might be used, for example for contacting the user, or through participation in surveys and feedback. iHEARu-PLAY stores the usernames and e–mail addresses in such a way that only selected employees, in detail researchers of the institute working on the iHEARu project[2] (Schuller et al., 2014b), have access to it. This assignment will only be used to identify a user's data, if at a later time, the user likes his or her data to be deleted from the database. Data collection is always in accordance with applicable data protection regulations. The aggregated data will not be used to personally identify a user on the mentioned purposes.

Annotations: Annotations are collected using a smartphone application or a standard PC and can be done any time and anywhere as long as audio can be played back to the user. Even though iHEARu-PLAY's primarily intended area of use is the labelling of audio databases, it is basically modality-independent, i. e., images and videos can also be imported.

Speech Recordings: Speech data is also collected using a smartphone application or a standard PC, which will allow the user to record prompted voice messages and upload them to the iHEARu-PLAY server – of course, only if the user's explicit consent has been given. Microphones which are embedded in most laptop PCs, tablets, and smartphone devices can be used to perform the recordings. With this feature, collection of speech data under real-life conditions in the wild (e. g., different microphone types, devices, background noises, reverberations, etc.) of a large number of subjects with different geographic origins, languages, dialects, cultural backgrounds, age groups, etc. will be possible. Those speech signals collected in the wild may also contain different types of surrounding noises such as crowd noises from events, traffic noises, and other city noises.

[2] http://www.ihearu.eu/

3.1.2. Data Storage

In cases where personal information is entailed in some scenario, there should be complete guarantee that the delivered, stored, and transmitted data are managed only by the administrators with the appropriate access permissions. State-of-the-art technologies for secure storage in a locked server, delivery, and access of data will be used. Firewalls, network security, encryption, and authentication will be used to protect the collected data.

3.1.3. Data Access and Usage

All given voluntarily information, annotations, or speech data that a user creates are automatically saved by iHEARu-PLAY and internally connected to the user's account. Therefore, all given data will be mapped to a pseudonym, which is internally also mapped to the username and e–mail address. All data will be stored electronically, always in an anonymised or pseudonymised form, and used exclusively for scientific research purposes; this means in particular:

- Access to a user's username and e–mail address is restricted to the selected employees of the service.
- The information which maps the username and e–mail address to the related generated and given data will at no time be shared.
- The given user's pseudonym and its linked pseudonymised data as well as the anonymised metadata will be shared with third party research bodies within and outside the EU only on a license base.
- Metadata, annotations, and recordings will be stored after the end of the iHEARu project (Schuller et al., 2014b) for use in follow up research projects. This will greatly help follow up research, ensure reproducibility of results, and eliminate the need to record new data over and over.
- Randomly selected samples of the pseudonymised audio data will be played to volunteers for annotation or for perception studies, either at a lab, or through crowdsourcing websites on the internet like iHEARu-PLAY.
- Samples of the audio data, generated figures, and anonymised metadata for statistical demonstrations can be used in scientific or public presentations. These figures can also be used in online scientific or public publications.

3.1.4. Data Changes or Deletion

The user has always the right to learn about all of his or her given information stored in the database, to correct it, or have them deleted. As a user, you can also access the relevant personal data at any time, change it, or remove it on your own. If the user disagrees with this general privacy policies, or wants to use the services of iHEARu-PLAY no

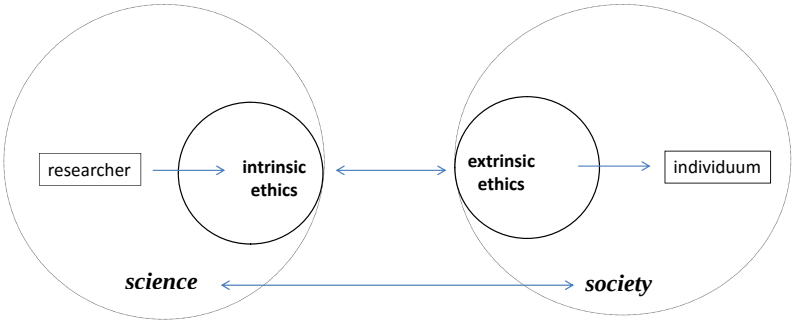

Figure 4: Intrinsic ethics (science) vs. extrinsic ethics (society).

longer, the user can request a deletion of the user account at any time. In the event of termination of the membership or a blocking of the account, iHEARu-PLAY will delete the e–mail address and the username of the user. Users always have the right to request the deletion of own annotations or recorded data through iHEARu-PLAY, and to obtain information about with whom the data have been shared. Nevertheless, data that have been already shared with third parties or that have been already used for publications cannot be deleted post hoc.

4. General Considerations

In this Section, we will broaden the view and present some general considerations on the role of science and an individual researcher (as part of science) on the one hand and society and individual (as part of society) on the other hand. In Figure 4, this interplay is depicted schematically. We tell apart intrinsic from extrinsic ethics, cf. (Batliner and Schuller, 2014): "In short, intrinsic ethics aims at producing sound scientific results; extrinsic ethics aims at the societal requirements that scientific results have to meet.".

Following the rules of intrinsic ethics or breaking them has both impact on society in general and – when an individual is directly involved – on the individual in particular. Intrinsic ethics pertains all aspects and criteria that have to be taken into account for producing 'good' science: That is what we learn in introductory courses, both at the beginner and at the post-graduate level, what we can read about in discussions in scientific journals; eventually, misconducts can lead to a public debate in newspapers and governmental bodies. Catchwords are: No plagiarism, sound reasoning (a somehow vague and generic, but very important requirement), adequate experimental design, adequate analysis and evaluation, correct use of (inferential) statistics (null hypothesis testing), adequate interpretation of results and taking into account possible impact, and last but definitely not least, adequate presentation of results to the public – for instance, by using 'common language measures' that can be conveyed easily to the non-expert (McGraw and Wong, 1992).

Possible impact leads to extrinsic ethics, where privacy considerations are in the fore for any study that employs individual subjects (or, in the case of big data, information that might be traced back to individual subjects); above, we have detailed our approach within the iHEARu-PLAY game. Furthermore, which consequences it will have when we transfer results onto real life – for society in general and for some individuals in particular; think of the impact of a wrong modelling on therapy such as proof of concept of new drugs in humans with possibly detrimental consequences. Other examples of a direct impact on an individual is a wrong therapy based on faulty classification and subsequent modelling (recognition/generation/teaching of states such as emotions in the therapy of children with Autism Spectrum Condition, cf. Schuller et al. (2014a)), or sarcasm/emotion detection in conversations with automatic call-center agents. Besides such direct impact on the individual, there is indirect impact as well: misleading financing which prevents financing of promising approaches, and wrong societal decisions with unfavourable consequences for the individual.

Summing up, science (and every individual researcher) has an obligation to provide meaningful results – if it only were in exchange for the money given from society (public bodies, etc.). Now, we can as well turn the tables: "don't ask what science can do for you – ask what you can do for science.". Provided that science really creates not only meaningful but also useful results, and especially in the case of important goals (for instance, diagnosis, and therapy of diseases such as speech pathology or autism), society should support science, and this means any individuals belonging to society as well. Naturally enough, this cannot be based on an obligation to deliver (such as taking part in experiments) but on the same terms as people are invited to donate blood – on a voluntary basis, with some incentives. This can be some payment for taking part in the experiments, credits for students, or – as we have illustrated above – simply fun in playing iHEARu-PLAY, while getting rewarded with prizes and scores and at the same time supporting science.

5. Conclusions and Outlook

We have shown ethic considerations concerning iHEARu-PLAY, a modular, cross platform, browser-based crowdsourcing game for collecting large-scale, real-life data for advanced paralinguistic tasks. When dealing with crowdsourcing, different ethical and privacy issues arise, e. g., concerning ethics in recording of private content, data col-

lection, annotation of crowdsourced data, and storage of the data. As the user will reveal personal data to iHEARu-PLAY, it is explained in detail what information and data iHEARu-PLAY collects, what usage the data has for us as researchers, with whom the data will be shared, and what will be done to protect a user's privacy. We also specified the measures taken to guarantee privacy in more detail. Moreover, we addressed general considerations on the role of science and researcher on the one hand (both have an obligation to provide results and data that have been financed by society), and society and individuals on the other hand (both should provide resources especially for important social issues). iHEARu-PLAY is still being developed; due to its modular architecture, rapid addition of new features is possible and planned. Besides integrating different kinds of features, our future work will focus on improving the quality management of the generated labels and recordings.

6. Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement No. 338164 (ERC Starting Grant iHEARu). We thank audEERING GmbH for technical support and sponsoring lottery prizes.

7. References

Adda, G. and Mariani, J. (2013). Economic, Legal, and Ethical Analysis of Crowdsourcing for Speech Processing. In M. Eskénazi, et al., editors, *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription, and Assessment*, pages 303–334. Wiley, Chichester.

Adda, G., Besacier, L., Couillault, A., Fort, K., Mariani, J., and De Mazancourt, H. (2014). "Where the data are coming from?" Ethics, crowdsourcing and traceability for Big Data in Human Language Technology. In *Crowdsourcing and human computation multidisciplinary workshop*, Paris, France, September. CNRS.

Batliner, A. and Schuller, B. (2014). More Than Fifty Years of Speech Processing - The Rise of Computational Paralinguistics and Ethical Demands. In Cerna, editor, *Proceedings of ETHICOMP 2014*. 11 pages.

Burkhardt, F., Eckert, M., Johannsen, W., and Stegmann, J. (2010). A database of age and gender annotated telephone speech. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Eskenazi, M., Levow, G.-A., Meng, H., Parent, G., and Suendermann, D. (2013). *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*. Wiley, Chichester. 356 pages.

Foundation, D. S. (2014 (Version 1.7)). Django web framework. Computer Software.

Good, B. M. and Su, A. I. (2011). Games with a scientific purpose. *Genome Biology*, 12(12):135.

Hantke, S., Eyben, F., Appel, T., and Schuller, B. (2015). iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing. In *Proc. 1st International Workshop on Automatic Sentiment Analysis in the Wild (WASA 2015) held in conjunction with the 6th biannual Conference on Affective Computing and Intelligent Interaction (ACII 2015)*, pages 891–897, Xi'an, P. R. China, September. AAAC.

McGraw, K. and Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, 111:361–365.

Schmidt, F. A. (2013). The Good, the Bad and the Ugly: Why Crowdsourcing Needs Ethics. In *2013 IEEE Third International Conference on Cloud and Green Computing*, pages 531–535, Karlsruhe, Germany.

Schuller, B. and Batliner, A. (2014). *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley.

Schuller, B., Marchi, E., Baron-Cohen, S., O'Reilly, H., Pigat, D., Robinson, P., Davies, I., Golan, O., Fridenson, S., Tal, S., Newman, S., Meir, N., Shillo, R., Camurri, A., Piana, S., Staglianò, A., Bölte, S., Lundqvist, D., Berggren, S., Baranger, A., and Sullings, N. (2014a). The state of play of ASC-Inclusion: An Integrated Internet-Based Environment for Social Inclusion of Children with Autism Spectrum Conditions. In Lucas Paletta, et al., editors, *Proceedings 2nd International Workshop on Digital Games for Empowerment and Inclusion (IDGEI 2014)*, Haifa, Israel, February. ACM, ACM. 8 pages, held in conjunction with the 19th International Conference on Intelligent User Interfaces (IUI 2014).

Schuller, B., Zhang, Y., Eyben, F., and Weninger, F. (2014b). Intelligent systems' Holistic Evolving Analysis of Real-life Universal speaker characteristics. In Björn Schuller, et al., editors, *Proceedings of the 5th International Workshop on Emotion Social Signals, Sentiment & Linked Open Data (ES^{3}LOD 2014), satellite of the 9th Language Resources and Evaluation Conference (LREC 2014)*, pages 14–20, Reykjavik, Iceland. ELRA.

Seedhouse, D. (1998). Against medical ethics: a response to Cassell. *Journal of Medical Ethics*, 24(1):13–17.

Silberman, M. S., Irani, L., and Ross, J. (2010). Ethics and tactics of professional crowdwork. *XRDS*, 17:39–43.

Turk, A. M. (2016). Amazon mechanical turk participation agreement, 15.01.2016. URL: *https://www.mturk.com/mturk/conditionsofuse*.

Zhai, H., Lingren, T., Deleger, L., Li, Q., Kaiser, M., Stoutenborough, L., and Solti, I. (2013). Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. *J Med Internet Res*, 15(4):e73, Apr.

Liability Specification in Robotics

Ethical and Legal Transversal Regards

Agnes Delaborde[1,2], Noémie Enser[2], Alexandra Bensamoun[2],Laurence Devillers[1,3]

[1]LIMSI-CNRS, Université Paris-Saclay, Orsay, France

[2] Centre d'Études et de Recherche en Droit de l'Immatériel, Université Paris-Saclay, Sceaux, France

[3] Université Paris-Sorbonne IV, Paris, France

E-mail: agdelabo@limsi.fr, noemie.enser@u-psud.fr, alexandra.bensamoun@u-psud.fr, devil@limsi.fr

Abstract

Beyond the implementation of moral considerations, an ethical robot should be designed in a way that foresees the potential damages it could cause, and that also anticipates the way the human beings in its environment (from the designer to the user) could be held responsible for its acts. In this present study, the authors offer to consider the actions of the robot under the French liability regime for the actions of things. In these conditions, the designer, the manufacturer and the user could be held liable for the actions of the robot. As a preventive measure, the robot would be subject to a mandatory insurance assumed by the user, and the designer would endow the interface with a log of data that could be assessed by an expert. As part of a curative approach, the authors present realistic case studies in which the robot could cause damage, with a determination of the liability based upon the liability regime for the action of things.

Keywords: Roboethics, Law, Robot liability

1. Introduction

Although it is commonly agreed that the robot cannot be considered a legal person in similar terms as the human being, its autonomy raises the question of its legal liability. Indeed, the actions of an autonomous robot can lead to damages, which need to be both anticipated and compensated for. A vast amount of research works has addressed the question of the liability of the robot, trying to determine which entity would be held responsible for the actions of the robot.

First and foremost, the authors wish to note that the term "action" does not imply any consciousness or intentionality in the robot: the actions of the robot are only driven by the commands of its algorithms. The "autonomy" of the robot refers to the fact that it is endowed with the capability of selecting the actions to perform, through a capture of the environment from sensors and a decision-making associated to this capture, without having to be manually operated by the user.

This research work takes place in the context of the French project TE2R "**T**races, **E**xplications et **R**esponsabilités du **R**obot"[1] ("Footprints, explanation and liability of the robot"). This project features a close collaboration between researchers in law and robotics, and aims at analyzing in which way the behavior of the robot can be tracked and explained, so as to facilitate a search for liability. The project will result in proposals for the design of ethical robotic interfaces. As (Asaro, 2011) points out in his study on the legal issues raised by robotics, there is often an overlap between what is considered moral, and the issues dealt with by the law. Indeed, the ethical nature of a robot does not only rely on it following moral and common sense rules; it also means that the robot should be designed in a way that foresees the potential damages it could cause, and in a way that also anticipates the way the human beings in its environment (from the designer to the user) could be held responsible for its acts.

In this present study, the authors look into the way the manufacturer, the designer and the user could each take responsibility in the actions of the robot, and how subsequent potential damages caused by the robot could be preventively and curatively dealt with. Therefore, a new liability regime for the actions of things could be established for robots, based primarily on a mandatory insurance assumed by the user, on the model of the existing insurance system for motor-driven land vehicles. Indeed, trying to forestall and limit the damages a robot could cause cannot always prove to be sufficient: damages will occasionally happen. Should an incident occur, a mandatory insurance would reveal its importance, since the collectivization of risks will imply that the insurer assumes the duty of repairing damage, thus facilitating the compensation.

The authors' approach could allow anticipating and facilitating a legal search for liability. The subject is broached from the point of view of French law, according to the authors' domain of expertise, but is thoroughly explained so as to potentially lead to discussions and adaptations with researchers from foreign legal traditions. Indeed, one of the objectives of the authors' work consists in developing the subject on an international level.

The liability linked to the actions of the robot depends essentially on its legal status. In consequence, the authors will offer firstly an overview of the possible legal status of the robot. In a second part, the authors will define the applicative domain of this study, which addresses exclusively the socially interactive robots for specific tasks. Next, propositions to integrate preventive and curative approaches in a robotic system will be presented. In that section, the authors will, with the support of a case study, elaborate a pattern for the distribution of the liabilities, and check this pattern against several other case studies. In a final point, the authors will discuss the

[1] Interdisciplinary project LIDEX Paris-Saclay and Institut Société Numérique

way this study could be expanded in the light of different disciplinary fields involved in the design of robotic interfaces.

2. State of the art: a legal framework for the robot

2.1 An international motivation

Legal rights of robots have been looked into ever since the rising development of roboethics. (Asaro, 2007) offers an overview of legal concepts that can be applied to solve certain ethical issues in robotics, for example by considering the robot on the same status with corporations. (Pagallo, 2010) addresses the bond between the owner and the robot, through an analogy with Roman masters and slaves where the master would be liable for the activities of his robot, highlighting the fact that such a system may not be fully applicable to face legal responsibility and ethical issues. (Calverley, 2008) explores in which terms a non-biological machine could be theoretically considered a legal person, according mainly to its degree of intentionality and autonomy. In the framework of the European Robolaw project, that aims for a socio-economic regulation in terms of ethics and law, (Bertolini and Palmerini, 2014) mention notably the possibility of granting the robot with a legal personhood, and of using "black boxes" to identify the decisions of the robot.

In a more general context, the international community easily admits that pluridisciplinary collaboration is compulsory so as to legislate for new technologies. (Askland, 2011) highlights notably that law cannot adapt correctly without a deep understanding of technical functioning and constraints. For example, the type of algorithms used in the systems impacts the liability determination: (Matthias, 2004) details why the outputs of systems based on neural networks and genetic algorithms cannot exactly be predicted, which thus prevents determining whom could be held liable, among the manufacturer, the programmer, and the operator. The issue raised in that study is still relevant.

Obviously, the process of regulating new technologies does not only involve jurists (or policymakers) and the creators of the technologies, but it can also concern any field of research, such as medicine, philosophy, economics. For example, (Ludlov et al., 2015), in their study on the relevance of resorting to existing regulatory regimes to assess new and emerging technologies, highlight the fact that anticipating the risks of future technologies could be broached from an ethical point of view, as well as societal, medical, environmental or safety point of view, and that each factor should not be assessed in isolation.

According to literature on regulating new technologies, the importance of pluri- and transdisciplinarity is widely stressed, and the authors of the present research particularly agree on that point.

As for the legal framework of the robot, studies on law and robotics show that its legal status is still to be defined. Indeed, the definition of this status is as of now only prospective, since robots are not yet widely distributed, and the complexity and novelty of the technologies involved in robotic systems does not make the regulation easier.

2.2 French positive law[2]

The first thing to note is that the French law is built upon a *summa divisio* between "persons" and "things": any entity belongs either to one category, or the other. The principle relies on an exclusion, which states that only the "person" category is restrictive; from a legal point of view, anything that is not a person is necessarily a "thing".

At the present time, the French positive law universally recognizes only two subcategories of persons: natural persons (the human), and juridical persons (entity that is not a natural person, authorized by law, with duties and rights, recognized as a legal authority, having a distinct identity, a legal personality, such as a corporation).

The law considers that everything outside these two subcategories of persons is a thing. Things can by no means be subjects of rights, in that they cannot be entitled to have duties or rights. They can only be objects of rights, that is to say objects on which a person can exercise their rights and duties.

In the field of liability, the regime differs whether the damage is caused by a person or a thing: the French law distinguishes the person's liability (for their own action or under vicarious liability) or the liability for the action of a thing.

The idea for a liability regime for the actions of things emerged during the industrial revolution, when a regime based solely on the liability for a person's actions proved to be insufficient. Indeed, the increasing number of occupational accidents led to a global reflection, since in many circumstances no person seemed to be involved in the damage occurrence. Judges, then the legislator, intervened to create new mechanisms so as to render a person liable not only for their own action, but also for the action of a thing under their control.

This liability regime does not require that the owner has committed negligence: it is necessary, and sufficient, that a thing causes damage for its custodian to be held liable. The custodian of the thing is the one who has effective powers of use, direction and control over it.

The general liability regime for the actions of things became rapidly insufficient, and the French legislator introduced more specific regimes. This is why, beyond the common regime that applies to all things, the law admits the specificity of certain types of things, and establishes a particular regime for them, as is the case with motor-driven land vehicles and animals. However, a robot

[2]Positive law: statutory man-made law currently in force

could not fit into one of these two categories: it is endowed with a level of autonomy that most of the things lack, and, contrarily to an animal, of which the owner is deemed liable for its misdeed, a robot is *designed* and *used* by a human being.

For these reasons, the authors consider that the law could adapt: perhaps not radically by establishing a judicial personality for the robot, but by taking into account the specific nature of the robotic thing.

The reflection carried out in the framework of the authors' project TE2R rests notably upon an initiative from the CERDI laboratory on the subject of law and robotics, resulting in a collaborative book in French (Bensamoun et al., 2016).

3. Applicative domain

In this present study, the authors will only consider robots intended for the social communication with humans, the development of which currently representing an extensively active field of research. This specific type of robot will be specified in this section, and its applicative domain will be defined in terms of task and end-user characteristics.

3.1 Definition of the domain

3.1.1 Nature of the robot and robotic tasks

The type of robots concerned in this study meets three cumulative points.

First, the type of robots falls within the category of "socially interactive robots", as defined notably in (Fong et al., 2003). In other terms, the propositions would apply to robotic interfaces endowed with the ability to engage a social dialogue (whatever its complexity) with the end-user through natural human communicative media (e.g. speech, gestures, or any human-understandable codification).

Secondly, the robot should be autonomous, which implies that it is not teleoperated by a human being and that it is endowed with decision-making mechanisms (falling within the Artificial Intelligence domain). The robot may have to resort to the human's advice for specific situations, and act accordingly to his/her proposition. Nonetheless, the decision to ask guidance from the human happens occasionally and is an initiative of the robot itself. The authors also wish to note that, although the robot is expected to obey to the user's orders, this cannot be considered on the level of autonomy: obeying to orders is part of the expected features of the robot, and modules involved in the achievement of the task asked from the user are driven autonomously.

Finally, the abilities of the robot should fall into, alternatively, one of these fields: gaming (play a game against/with the user), personal assistance (e.g. schedule management, office automation), or social dialogue (e.g. small talk, story-telling).

3.1.2 Software and hardware

This study concerns both the software and hardware elements of the robot. The notion of software applies to the algorithms and methods related to the artificial intelligence domain, directly involved in the interaction with the user and the performing of the tasks expected from the robot (dialogue managers for communicating, navigation so as to maneuver in the users' environment, etc.); the hardware concerns the physical elements of the robot that are involved in the tasks that it can perform for the user.

The exhaustive definition of what falls within the category of "software" or within the category of "hardware" is not a task the authors intend to perform in this study. Indeed, they are aware that there are different levels of programming in any electronic device (integrated circuits, motors, etc.), which are part of the overall capacity of the device to adapt to its environment. A fault due to the programming of such elements could lead to damages, but this fine distinction will not be considered by the authors, but left as open questions addressed to the community.

3.1.3 The end-user

The end-user concerned by this study uses a finished product, for which he or she does not have the knowledge or the will to make changes.

3.2 Exclusion

Industrial robots and household appliances are not taken into account in this study, as long as they are not aimed at an intelligent social interaction with the user (see 3.1.1).

For this present study, the authors also opt not to consider the medical domain (healthcare robots for impaired users, or nurse-robot-user triad), since this domain implies deontological considerations and is relative to specific legal aspects.

Once all the conditions listed in this section are met, the authors offer two approaches. The first approach consists in anticipating the problems related to the liability of the robot; the second approach tries to concretely settle the issues relative to a search for liability, in the case of damage.

4. Preventive approach

In the authors' approach, preventive measures can be established so as to facilitate a search for liability. These measures shall not be understood as propositions meant to reduce the risks of damage, but to facilitate a search for liability (and thus indirectly make the process of damage repair easier).

In the first part of this section, the authors briefly look into a design of the robotic logs which could allow the diffusion of relevant information. They also assess the possible legal and ethical relevance of such data. Implementing the possibility for the robot to keep and

broadcast specific logs represents a step that can be taken at the stage of the design of the robot (during the step of "pre-commercialization"). These logs could thereafter be assessed during audits performed in the context of a mandatory insurance ("post commercialization"). This p-oint will be detailed in the second part of this section.

4.1 Pre-commercialization

The robot could be endowed with a data-logging system that would allow experts to track the data captured by the robot and the decision it has made.

The nature of the data that needs to be made available in the logs of the robot remains a research question that the authors look into. However, they define three types of data in the context of a robot interacting socially with a human: A) environment-related data, B) decision-making data and C) decisions.

The data concerned by the case A represents the information the robot has captured in its environment, and that can prove to be useful for the task expected from the robot. One can consider finding in this data, for example, information about the user (e.g. his/her name, his interactional preferences), about the surroundings of the robot (e.g. localization of the rooms in the user's house, of objects), or data about the internal state of the robot (physical aspects: position of its joints, or simulation of cognitive mechanisms: mood, social attitude). This information would allow knowing the global context of the robot (what it was doing at a specific time, where, with whom, etc.). In a robotic system endowed with emotion detection from speech (Devillers et al., 2015), for example, the detected paralinguistic data can be added to the logs for each speaker turn of the user.

The decision-making data (case B) represents the input data that take part in the decision process. This data will be a subsection of the data detailed in case A, that is specifically used to assess the rules of the decision-making system. For each decision it has made, the robot would be able to tell the nature of this decision (case C), such as "ask the user about his health", or "update the planning", and also tell in function of which elements this decision was made (user's demand, change in the environment, etc.).

The algorithms and models that take part in the decision process can obviously not be made available, since they would usually fall within the industrial secret. Nonetheless, the authors consider the possibility that the input and output data could allow a transparency of the reasoning performed by the robot, while still preserving a "black-box"-like structure.This point, along with a description of the nature of the logs, is addressed more specifically by the authors in other current research works.

From a legal point of view, logs have already been produced in French courts (browser logs). According to French constant jurisprudence, logs have a probative value close to the value of testimonies.

The reliability of the logs would rely on the fact that the user (and non authorized individuals) cannot edit them. The processing chain of the logs would thus need to be transparent so as to preserve their evidential force, hence the necessity to ensure the phase of the creation of the logs, and the data back-up process. The creation would be in the hands of the designer, who would decide the log contents according to the types of data that would be deemed relevant for a search of liability. The back-up could either be dealt with by the designer or by the manufacturer, which would ensure that the data cannot be modified (internal non-editable memory of the robot, secure remote log server).

Since the robot will be used in the end-user's environment, the biggest part of the stored data will concern the user. This point naturally raises issues in terms of privacy, which could be a major obstacle in the development of such a proposition. The authors are fully aware of the implications of keeping logs about an individual, and one of their concerns consists in determining a legal, technical and ethical compromise between what could be done, and what should be done.

The authors consider that data-logging, despite its obvious usefulness, could be replaced by simulations. These simulations would consist in placing the robot in specific situations that could allow testing several aspects of the robot, without giving explicit access to potentially private data. One straightforward example would be that, if the expert needs to check the robot is physically safe, he/she can, for instance, place a finger on a joint and check whether the robot pinches it or not. The assessment can get more complicated if the expert needs to check, for example, that the robot has not been used for unlawful activities (or improper use). First, exhaustively testing could be tedious and uncertain, and the definition of a global checking protocol would be almost impossible, since it is intimately dependent on the robot's capacities. Also, checking that the robot has not learnt unlawful or risky behaviors (if the robot is endowed with learning algorithms) would require placing it in a situation that could possibly trigger an improper behavior. However, algorithms involved in the behavioral decisions of the robot are not necessarily determinist.

The definition of an adequate level of data-logging, and the relevance of considering simulations in lieu of (or along with) data-logging, are current research tasks of the authors.

4.2 Post commercialization

The preventive approach offered by the authors relies upon a mandatory insurance assumed by the end-user. Such mandatory insurance mechanisms exist in most countries, notably for cars. As with a person who wishes to use a car has to insure it, the owner of a robot would be required to contract insurance for it.

In the framework of this insurance, a regular and mandatory audit could be performed. This audit would

include a retrospective check to make sure that the functionalities of the robot have not been distorted (by analyzing the logs or through simulations, as defined previously), and a prospective check to assess that the current state of the robot will not cause subsequent damages.

Depending on the robots functionalities, the analysis of the log could either be presented through a direct reading of logs (which would imply either a certain amount of knowledge from the expert, or a thoroughly conceived log), or through oral closed-ended questions that could allow the expert to query the database of the robot for specific information. Placing the robot in specific situations would allow the expert to determine whether the actions of the robot are still safe.

The expert's validation could allow the user to continue to benefit from a bonus on his/her insurance premium, or, on the contrary, the absence of validation would lead to a penalty. Should the expert note severe defects, he/she could forward this information to competent authorities.

The notion of what needs to be checked during this audit will strongly depend on the tasks and capacities expected from the robot. The authors' future research work will notably consist in trying to define some general categories for the damages that could occur, and the way they could be retrospectively and prospectively checked.

5. Restorative approach

One must imagine that, despite the preventive approach, the robot could be the cause of damage.

French law knows two types of damage, which can absolutely add up to each other: material damage and non-material damage. Material damage can be quantifiable in pecuniary terms, that is to say either an impoverishment or a deprivation of a legitimately foreseeable increase in wealth. This is notably the case when an object gets ruined, or when the victim has to face significant medical costs following damage inflicted on his/her bodily integrity. On the other hand, the damage is considered non-material when it affects psychologically the victim (rather than his/her patrimony), for instance further to the destruction of an object with a high sentimental value. Another possible case of non-material damage would be if the victim cannot dedicate himself/herself to an appreciated activity anymore, due to damage on his/her bodily integrity. For both types of damage, the victim can legally expect to be granted adequate financial compensation.

However, the authors deem it noteworthy to point out that purely affective damage caused by a robot can, by no means, lead to legal compensation. This question can legitimately be raised with companion robots, which are expected to have an emotional impact on their owners. Nonetheless, damage resulting from the degradation of the affective relationship between the human and the robot does not constitute a type of damage for which the victim can ask reparation; indeed, the law does not deal with emotional issues.

In this section, the authors will first look into a case study that will lead their reflection upon the distribution of the liability in case of damage. Secondly, this pattern will be checked against several other study cases, so as to assess whether its application may be generalizable.

5.1 Development of a pattern of liability distribution

Depending on the origin of the dysfunction, three different persons could be held responsible: the user, the manufacturer or the designer of the system. So as to define the liability distribution, the authors reasoned from this case study: a companion robot must record its user's work schedule and register every professional appointment. In the considered case study, the robot failed to remind an important meeting, and the user loses his/her job.

The authors selected this scenario because it is totally realistic in regards to the current state of technology. One could argue that this specific example is not exactly a case that highlights the specificities of a robot (a tablet could manage a schedule just as well). However, endowing the robot with this capacity entails several technologies for the transmission of data with the user, since the interaction is oral, and robots are not all endowed with a screen (which could allow the user to set and check easily by his/herself the state of the schedule).

Besides, the robot is expected to accompany the user in his/her daily life, and has to manage its own activity prioritization. Contrary to a tablet application that would simply make a notification icon pop or start an audio alert (disregarding what the user is currently doing with his/her tablet), the robot is expected to behave socially with the user, which will imply that it cannot go beeping right in the middle of, for example, a small talk conversation with its owner. In this way, one can expect that an application as simple as a schedule manager will be nested in amore complex global behavior manager, and that it will rely upon more sophisticated communication functionalities than simply launching an application and clicking.

5.1.1 The user's liability

The user own responsibility could be held against him or her, if the dysfunction is due to a human error relative to the use of the robot. For example, if the user did not properly save his appointment, he cannot successfully seek the liability of a third party. This remains equally true, even if the user's mistake is due to poor ergonomics in the robot. Indeed, neither the manufacturer nor the designer can be legally held liable, if the interface is poorly designed. This aspect would result in trade sanctions (users would turn to better designed robots), but would in no way engage legal sanctions.

In a completely different context, the user's liability could be held if it is proven, through the audit or any other mean, that he misuses the robot, for example by using the robot

to perform punishable offences. In this specific case, it is not inconceivable that a designer may sue a user for improper use of his technology on the ground of a copyright infringement.

5.1.2 The manufacturer's liability

The manufacturer's liability could be sought every time the dysfunction is due to a hardware failure. In the previous case, the manufacturer could be held responsible, for example, if the memory of the robot got deleted due to an electronic failure, or if a material deficiency prevented the transmission of the information.

5.1.3 The designer's liability

The designer's liability could be sought every time the dysfunction is due to a failure in artificial intelligence. Using the same example, the designer could be held responsible if the memory of the robot got deleted because of a data storage issue, or if the robot forgot to remind the appointment because of a mismanagement in the tasks prioritization.

5.2 Assessment of the pattern relevance

Through different case studies, the authors will show that the pattern developed in section 5.1 may adapt to several situations.

The robot did not notice the presence of the user, and it accidentally knocks him/her down when moving. As a consequence, the user breaks his/her leg. The liability of the manufacturer could be engaged if it can be proven that the user was in a dead angle (e.g. the robot has no possibility to see or feel the user in this position), since reducing the extent of dead angles (thus making the device safer) is at the charge of the manufacturer. The manufacturer could also be held liable if the sensors present a fault, as long as the user is not the direct cause of this fault (improper use). The designer could be held liable if the decision made by the robot, based on correct data transmitted by the sensors, led it to move in the wrong way.

In another case study, one of the tasks expected from the robot consists in detecting if a stranger enters the user's home. When this case occurs, the robot has to ask the user if there is a problem, and if the answer is positive it broadcasts an alert to an external security company. In the present scenario, the robot detected a stranger, and asked for the user's advice, who confirmed there was a problem. The robot did not send alert, and the user has been attacked. In the case where the robot did not correctly understand the user's response, the manufacturer's liability could be engaged if the sensor presented a fault, while the designer would be held liable if the speech detection resulted in a misunderstanding. If the robot correctly understood the answer, but did not make the decision to call the security center, the artificial intelligence may be at fault (designer's liability), or the transmission system (manufacturer's liability).

6. Conclusion

To increase the acceptability of the robot, the latter has to follow the rules of the society in which it is being used. If moral considerations could be a step of implementation, including the robot in the legal framework is essential for its acceptation.

In the TE2R project, the authors look into the way the data saved by the robot, and the decision it makes from this data, could facilitate a search for liability. In this present study, the authors offer tracks to consider the robot under the light of the existing legal system, as an object that is autonomous in its decisions and actions, but that would still involve the responsibility of the human beings in its environment. They offer to consider the actions of the robot under the liability regime for the actions of things. They show that from a legal point of view, considering the robot holistically is not adequate to determine a liability regime. No matter how life-like and friendly the robot is, it still has no conscience or intentionality. The authors, nonetheless, admit that its capacity for autonomy could lead to the creation of a specific regime for the liability of things. In these conditions, the designer, the manufacturer and the user could be held liable for the actions of the robot. As a preventive measure, the robot would be subject to a mandatory insurance assumed by the user, and the designer would endow the interface with a log of data that could be assessed by an expert. As part of a curative approach, the authors present realistic case studies in which the robot could cause damage, with a determination of the liability based upon the liability regime for the action of things.

Establishing new legislations for robots that are meant to be used by consumers, often at their private domicile, raises many ethical issues. One major point that could be highlighted by this present paper would naturally be the issue of the privacy of the user's data, since the robot would be expected to keep a log of what happens at his/her place. In this respect, the authors explicitly agree that the definition of the legal framework of the robot must be a consensual and collaborative task. As previously mentioned, this work intends to lead to an international collaborative reflection, and in its present state voluntarily raises many highly interesting questions that will be addressed in the further development of the project, which will require feedbacks from experts from different research fields.

7. Acknowledgements

This research work is funded by the French Institut Société Numérique, LIDEX Paris-Saclay, in the framework of the project TE2R "Traces, Explications et Responsabilités du Robot".

8. Bibliographical References

Asaro, P. (2007).*Robots and Responsibility from a Legal Perspective*. In Proceedings of the IEEE Conference on

Robotics and Automation, Workshop on Roboethics, Rome, April 14, 2007.

Asaro, P. M. (2011). 11 *A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics*. Robot Ethics: The Ethical and Social Implications of Robotics, 169.

Askland, A. (2011). *Why law and ethics need to keep pace with emerging technologies*. In The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight. Marchant et al.(eds.), Springer. pp. xiii-xxvi.

Bertolini, A., & Palmerini, E. (2014). *Regulating Robotics: A Challenge for Europe*. Upcoming Issues of EU Law, 94-129.

Bensamoun, A. et al. (2016) *Les robots : Objets scientifiques, Objets de droits*. Under the supervision of Bensamoun A. Mare & Martin (Eds), Collection des Presses Universitaires de Sceaux. ISBN-10: 2849342157.

Calverley, D. J. (2008). *Imagining a non-biological machine as a legal person*. Ai & Society, 22(4), 523-537.

Devillers, L., Tahon, M., Sehili, M. A., & Delaborde, A. (2015). *Inference of human beings' emotional states from speech in human–robot interactions*. International Journal of Social Robotics, 7(4), 451-463.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003).*A survey of socially interactive robots*. Robotics and autonomous systems, 42(3), 143-166.

Ludlow, K., Bowman, D. M., Gatof, J., & Bennett, M. G. (2015). *Regulating Emerging and Future Technologies in the Present*. NanoEthics, 9(2), 151-163.

Matthias, A. (2004). *The responsibility gap: Ascribing responsibility for the actions of learning automata*. Ethics and information technology, 6(3), 175-183.

Pagallo, U. (2010). *The Human Master With a Modern Slave? Some Remarks on Robots, Ethics, and the Law*.In Proceedings of ETHICOMP 2010, Spain.