Ethical by Design: Ethics Best Practices for Natural Language Processing

Jochen L. Leidner and **Vassilis Plachouras**
Thomson Reuters, Research & Development,
30 South Colonnade, London E14 5EP, United Kingdom.
{jochen.leidner,vassilis.plachouras}@thomsonreuters.com

Abstract

Natural Language Processing (NLP) systems analyze and/or generate human language, typically on users' behalf. One natural and necessary question that needs to be addressed in this context, both in research projects and in production settings, is the question how *ethical* the work is, both regarding the process and its outcome.

Towards this end, we articulate a set of issues, propose a set of best practices, notably a process featuring an *ethics review board*, and sketch how they could be meaningfully applied. Our main argument is that ethical outcomes ought to be achieved *by design*, i.e. by following a process aligned by ethical values. We also offer some response options for those facing ethics issues.

While a number of previous works exist that discuss ethical issues, in particular around big data and machine learning, to the authors' knowledge this is the first account of NLP and ethics from the perspective of a principled *process*.

1 Introduction

Ethics, the part of practical philosophy concerned with all things normative (moral philosophy, answering the fundamental question of *how to live one's life*) permeates all aspects of human action. Applying it to Natural language Processing (NLP), we can ask the following core questions: 1. What ethical concerns exist in the realm of NLP? 2. How should these ethical issues be addressed? At the time of writing, automation using machine learning is making great practical progress, and this includes NLP tasks, but is by no means limited to it. As more areas in life are affected by these new technologies, the practical need for clarification of ethical implications increases; in other words, we have reached the level where the topic is no longer purely academic: we need to have solutions for what a driverless car should morally do in situations that can be described as ethical dilemmas, and in language and speech-enabled system ethical questions also arise (see below). Governments and NGOs are also trying to come to grips with what machine learning, which NLP also relies on, means for policy making (Armstrong, 2015).

In this paper, a more principled way to deal with ethical questions in NLP projects is proposed, which is inspired by previous work on the more narrowly confined space of privacy, which we attempt to generalize. In doing so, we want to make sure that common pitfalls such as compartmentalization (i.e., considering one area in isolation and solving problems in a way that creates problems elsewhere), do not hinder the pursuit of ethical NLP research and development, and we shall present some possible response options for those facing non-ethical situations to stimulate discussion.

Paper Plan. The rest of this paper is structured as follows: Sec. 2 introduces the concept of "ethical by design". After reviewing some related work in Sec. 3, Sec. 4 reviews ethics issues in NLP. Sec. 5 introduces a proposed process model and some possible responses for those facing ethics dilemmas. Sec. 6 discusses the shortcomings, and Sec. 7 summarizes and concludes this paper.

2 Ethical by Design

Ann Cavoukian (2009), a Canadian privacy and information officer, has devised a set of seven principles for *privacy by design*, a sub-set of which

Proceedings of the First Workshop on Ethics in Natural Language Processing, pages 8–18,
Valencia, Spain, April 4th, 2017.

we can generalize—so that they apply to general ethics standards instead of the single issue of privacy—as follows.

1. Proactive not reactive: by planning to do things in an ethical way we avoid having to react remedially to non-ethical situations more often than without a planning approach;
2. Ethical as the default setting: by making a commitment to pursuing originally ethical paths, we create alignment within organizations towards a more streamlined set of options that comply with common values;[1]
3. Ethics embedded into the process: a process firmly inclusive of ethics at all stages and levels is less likely to create accidental harm;
4. End-to-end ethics: ethics cannot be confined to a stage; it must be an all-encompassing property of a process from basic research over product design to dissemination or delivery, i.e. the full life-cycle of a technology;
5. Visibility and transparency: a process that is published can be scrutinized, criticized and ultimately improved by a caring community;
6. Respect for user values: whatever values a research institute, university or company may hold is one thing, being user-centric means to also consider the values of the user (of a component, product) and the subjects that take part in experiments (ratings, data annotations).

How could such principles be applied to NLP, concretely? We ought to make some practical proposals how to proceed e.g. in a research project or when developing a product to avoid ethical issues. To this end, we will now look at some potential issues, review best practices that are available, and then put it all together in the form of a process recommendation and possible responses for dealing with ethical issues as they arise.

3 Related Work

Prior work on the topics of ethics in NLP can be grouped into three categories. First, there is the general body of literature covering applied ethics and moral philosophy. Second, within computer science, there are discussions around big data, data mining and machine learning and their ethical implications, often focused on privacy and bias/discrimination. Few if any of these works have mentioned issues specific to language processing, but a lot of the unspecific issues also *do* apply to NLP.[2] Third, there is a body of works on professional ethics, often talked about in the context of curriculum design for computer science teaching (didactics of computing), governance and professional conduct and legal/ethical aspects of computing (computing as a profession, continued professional development).

Moral Philosophy & Ethics. We cannot even aspire to give a survey of centuries of moral philosophy in a few sentences; instead, we briefly sketch three exemplary schools of moral philosophy to represent the fact that there is no single school of thought that settles all moral questions.[3] Aristotle's "Nicomachean Ethics" (Aristotle, 1925; Aristotle, 1934)[4], Utilitarianism (Mill, 1879) and Kant's (1785) categorical imperative are just three examples of philosophical frameworks that can be used as a frame of reference to study ethics, including the ethics of NLP and its applications. Aristotle based his system on happiness (Greek εὐδαιμονία) as the highest attainable and ultimate goal for humans, and takes a consensus-informed view starting with those moral principles that people with "good upbringing" agree on. Kant's categorical imperative posits a decision criterion to decide whether an action is moral or not, namely whether we would want to lift up our behaviour so that it may become a law of nature. Utilitarianism suggests to maximise happiness for the largest number of people, which implies a quantification- and outcome-oriented aspect; however, it also contains a severe flaw: it can be used to justify unethical behavior towards minorities as long as a majority benefits.

Information & Big Data Ethics. There is a body of work within philosophy on information ethics (Allen et al., 2005; Bynum, 2008); big data has created its own challenges, which are begin-

[1] There is a catch, namely different people may agree to slightly different ethical premises, or they may draw different conclusions from the same premises.

[2] An edited collection on ethics and related topics in the context of artificial companions exists (Wilks, 2010), but as Masthoff (2011) points out NLP does not feature in it.

[3] For general background reading in ethics and moral philosophy, see Gensler (2011). For computing-related ethics background there already exist many suitable entries to the literature (Brey et al., 2009; Quinn, 2016; Stahl et al., 2016; Bynum, 2008; Bynum and Rogerson, 2004; Cary et al., 2003).

[4] named after its dedication to Nicomachus, likely either Artistotle's father or son

ning to be discussed. Pasquale (2015) provides a thorough analysis of the societal impact of data collection, user profiling, data vendors and buyers, and application algorithms and the associated issues; it contains numerous real case examples. However, the exposition does not appear to include examples likely to rely on NLP. Supervised learning, clustering, data mining and recommendation methods can account for the vast majority of examples (collaborative filtering, Apriori algorithm), which raises the questions of whether there will be a second wave of more sophisticated profiling attempts relying on NLP and neural networks.

Machine Learning & Bias. Since 2014, the Fairness, Accountability, and Transparency in Machine Learning (FATML, 2014) workshop series (originally organized by S. Barocas and M. Hardt at NIPS) have been concerned with technical solutions associated with a lack of accountability, transparency and fairness in machine learning models.

NLP Application Ethics. Thieltges, Schmidt and Hegelich (2016) discuss NLP chat-bots; in particular, they focus on the dilemma they call "devil's triangle", a tension between transparency, accuracy and robustness of any proposed automatic chat-bot detection classifier. Software that interacts with humans and/or acts on humans' behalf, such as robot control code or chat-bots will need to contain embodied decisions to ensure that the software acts *as if* it was a moral agent, in other words we would expect the software to act in a way such that a human acting in the same way would be considered morally acting (Allen et al., 2005). Most recently, Hovy and Spruit (2016) provided a broad account and thought-provoking call for investigation to the NLP community to explore their impact on society. To give an example of an unethical or at least highly questionable application, Mengelkamp, Rohmann and Schumann (2016) survey (but do not advocate) practices of credit rating agencies' use of social (user-generated) content, mostly unknown and unapproved by the creators of that data. Fairfield and Shtein (2014) analyze ethics from a journalism point of view, which bears some similarity with the perspective of automatic NLP, as journalists also scrutinize textual sources and produce text, albeit not algorithmically.

NLP Methodology Ethics. Fort, Adda and Cohen (2011) provide an early account of the ethical implications of crowdsourcing, the program-controlled automation of work conducted by anonymous human subjects.

Professional Ethics. Professional ethics is the integration of codification into education and continuous professional development, and the computing profession developed the *ACM Code of Ethics and Professional Conduct* (1992), which communicates a detailed set of 22 values; however, the compliance with them has been made voluntary, there are no negative consequences to people not adhering to them; further more, insufficient detail is given with regards to where moral boundaries are on specific issues, or how they may be obtained. The current code, then, falls short of an "algorithm how to be a good professional", if that can even exist. More recently (ACM, 2017), 7 principles were postulated to promote the transparency and accountability of algorithms: 1. Awareness; 2. Access and redress; 3. Accountability; 4. Explanation; 5. Data Provenance; 6. Auditability; and 7. Validation and Testing. The Association of Internet Researchers has published ethics guidelines (Markham and Buchanan, 2012), which have seen some adoption.[5] Perhaps the most interesting *empirical* finding to date is a survey by Fort and Couillault (2016), who posed ethics-related questions to French and international audiences, respectively, in two polls. For example, over 40% of respondents said they have refused to work on a project on ethical grounds.

This work draws heavily on Cavoukian (2009)'s proposal but goes beyond in that we propose a *process model* that makes intentional and non-intentional violations harder to go unnoticed. Our process is also informed by the holistic "Resources-Product-Target" (RPT) model of Floridi (2013). As he points out (Floridi, 2013, p. 20), many models focus rather narrowly on a "ethics of information resources", "ethics of information products" or "ethics of the informational environment" view (which he calls *microethical*). His counter-proposal, the RPT model (Figure 1), in contrast, is a *macro-ethical* approach that tries to avoid dilemmas and counterproductive effects by taking too narrow a view: RPT stands for "Resources-Product-Target", because Floridi's model considers the life cycle of producing an information product (output) from and information

[5] e.g. by the journal PeerJ. See also AoIR (2017 to appear) for an upcoming event on ethics in Internet research.

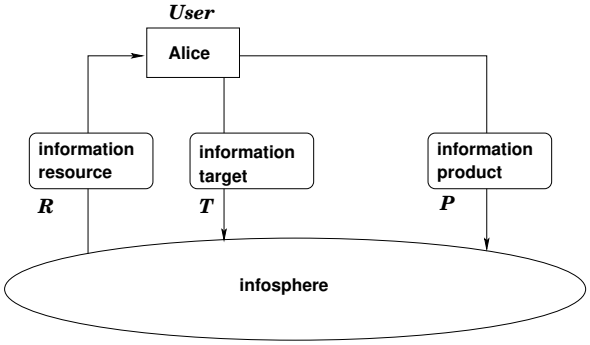

Figure 1: The Holistic Resources-Product-Target (RPT) Model after Floridi (2013).

resource (input), during which an effect on the environment (infosphere) is created (target). By considering the environment as well, compartmentalization, i.e. behaving ethically in a narrowly confined realm but not overall, can be avoided.[6] Information ethics within NLP is nascent, however there is a lot of general work that can be borrowed from, going back as far as early computer science itself (Wiener, 1954).

4 Ethical Issues in NLP

While the related work in the previous sections reviewed work on ethics including but not limited to ethics and NLP, in this section, we discuss the types of ethical issues that we are aware, and give some examples from the NLP literature. We will link each type to one or more parts of Floridi's model. An interesting question is what specifically is different in NLP with respect to ethics compared to other data-related topics or ethics in general. This question can be split into two parts: first, since it pertains to human language processing, and human language touches many parts of life, these areas also have an ethics dimension. For example, languages define linguistic communities, so inclusion and bias become relevant topics. Second, NLP is about processing by machine. This means that automation (and its impact on work) and errors (and their impact, whether intentional or not) become ethical topics. Furthermore, if NLP systems are used as (information) access mechanism, accessibility is another concern (inclusion of language-impaired users).

[6] A famous example of compartmentalization is the cruel dictator that is loving to his children at home. In an NLP context, an example could be a friendly and caring scientist that unwittingly abuses workers using a crowdsourcing API, because he needs gold data and has a small budget.

Unethical NLP Applications (pertains to P for Product in Floridi's RPT model). The earliest ethical issue involving NLP that the authors could find during the research for this paper surrounds the UNIX **spell(1)** command. Spell is a spell-checker: it prints out words not found in a lexicon so they can be corrected. However, in the course of its invocation, McIllroy's 1978 version (unlike Johnson's original implementation) emailed words that are not found in its lexicon to its implementer to support lexicon improvements (Bentley, 1986, p. 144); while this is technically commendable (perhaps even one of the earliest examples of log-file analysis), from a privacy point of view the author of a document may disapprove of this.[7] Youyou, Kosinski and Stillwell (2015) show that automated psychometrics—they use social media "like"s—now rivals human determination of personality traits; one interesting moral aspect is that when subjects wrote a piece of text they were likely not aware that in the future this may be possible, in the same way that many people who uploaded their photos online were not aware that one day face recognition at scale would reach maturity, which has now happened. In a similar spirit, Kosinski, Stillwell and Graepel (2013) demonstrate that other private personal traits and attributes can be computed from a user's data, including from their network of personal relationships. Thieltges, Schmidt and Hegelich (2016) describe another issue, namely that of chat-bots, which may act in a political way, such as steering or influencing a discussion or, worse, completely destroying meaningful human discourse by injecting noise: on Twitter, chat-bots made real conversation impossible for the topic channel `#YaMeCance` dedicated to reducing violence and corruption in Mexico, and real-life human activists were reportedly followed and threatened. It seems prudent that any bot self-identify as an automated entity, and from an ethical—if not legal—point of view, a respectful, low-traffic interaction is warranted. NLP developers should not participate in such efforts, not let themselves be instrumentalized by state actors or commercial interests, should withdraw from dubious projects, and pub-

[7] The authors of this paper do not know if all users of **spell(1)** were privy to this feature (we have not received a response from Prof. McIllroy to an email request for clarification while this paper was under review). In any case, it should be clear that the works cited here are listed to ignite the ethics discussion, not to criticize individual works or authors, whose work we greatly respect.

licize and disclose immoral practices.[8] In contrast, research into the automatic detection of such bots seems ethical, to increase transparency and reduce manipulation in a society, and this may require manipulative bots to be developed for test purposes; however, they should not be deployed, but be kept in sandboxed environments or just be implemented as simulations. Hovy and

Spruit (2016) point out the dual nature of some work: like a knife can be used to cut bread or to harm others, NLP may "have dual" use potential. There are two possible responses: either object if the non-ethical use is clearly involved in a project or product. Or alternatively, act conservatively and avoid obvious dual-use technologies entirely in favor of ethical-only use technologies (e.g. work on health-care applications instead of command-and-control software). Building an NLP application, like any other human activity that is a means to an end, can have an ethical end or not. For example, an NLP application could be an instance of an unethical application if its purpose is not consistent with ethical norms. If one adopts cherishing human life as an absolute moral value, developing a smart weapon using voice control would be an example of an application that is ethically wrong.

Davis and Patterson (2012) list identity, privacy, ownership and reputation as the four core areas of big data ethics. What is the range of potential ethical issues in NLP in specific? This paper cannot provide an exhaustive list, but we will try to give a nucleus list that serves to illustrate the breadth of topics that can be affected.

Privacy (pertains to T for target in Floridi's RPT model). Collecting linguistic data may lead to ethical questions around *privacy*[9]: Corpora such as the British National Corpus, the Collins COBUILD corpus or the Penn Treebank contain names of individuals and often substantial personal information about them; e-mail corpora to study the language of email (Klimt and Yang, 2004), or corpora of suicide notes and other sensitive psychiatric material (Pestian et al., 2012; Brew, 2016) constructed to study causes for terminating one's life are much more private still. Is the ability to construct a classifier that detects how "serious" a suicide note should be taken a good thing? It may prevent harm by directing scarce health resources in better ways, but does that justify the privacy invasion of anyone's personal good-byes, without their consent? Althoff, Clark and Leskovec (2016) describe an analysis of counseling conversations using NLP methods; here, perhaps because the patients are still alive, even stronger privacy protection is indicated. Another privacy-related issue is excessive government surveillance, which can lead to self-censoring and ultimately undermine democracy (Penney, 2016).

Fairness, Bias & Discrimination (pertains to T for target in Floridi's RPT model). Picture a spoken dialog system that is easy to use for a young male financial professional user with a London English pronunciation, but that may barely work for an elderly lady from Uddingston (near Glasgow, Scotland). As automated information systems are becoming more pervasive, they may eventually substitute human information kiosks for cost reasons, and then out-of-sample user groups could be excluded and left behind without an alternative. The internal functioning of NLP systems can raise questions of *transparency & accountability*: what if a parser does not work for particular types of inputs, and the developer does not communicate this aspect to an application developer, who wants to build a medical application that uses it. It is responsible behavior to disclose limitations of a system to its users, and NLP systems are no exception. In the context of machine learning, governments have started looking into the impact of ML on society, the need for policy guidance and regulation (Armstrong, 2015).

Abstraction & Compartmentalization (pertains to all parts of Floridi's RPT model). As mentioned earlier, Floridi's (2013) model was explicitly designed to overcome an overly narrow focus only on the input or project output. Abstracting over humans happens in crowdsourcing (see above) when work is farmed out to an API, which has individual humans behind it, but this fact can be all to easily ignored by the API's caller. If abstraction can lead to ethical ignorance in one dimension, compartmentalization can lead to the

[8] One reviewer has called our call for non-participation in un-ethical work "naïve"; however, we believe individuals can effect positive change through their personal choices, and especially in sought-after professions no-one has an excuse that they had to feed their family (or whatever justification one may bring forth). Also, by buying from and working for, more ethical companies, a pull towards increasing ethical behavior overall may be generated.

[9] Privacy as a concept is discussed in Westin (1967); See the recent seminal literature on big data privacy (Lane et al., 2014; Zimmer, 2010) for more in-depth discussions of data and privacy.

same in another. For example, an information extraction system that gets build without looking at the political context in which it is likely deployed may lead to unethical actions by a team of otherwise morality-oriented researchers.

Complexity (pertains to T as in target in Floridi's RPT model). Today's big data systems are cloud-based pipelines of interconnected data feeds and "black box" processes (Pasquale, 2015) combining and transforming a multitude of sources each, and these transcend individual organizational boundaries. This means that an organization offering e.g. an NLP API ceases control of how it may be used externally; this creates complex, hard-to-understand macro-ecosystems.

Un-ethical Research Methods (pertains to R as in Resource and T as in target in RPT). Doing research itself can be done in more or less ethical ways, so the discussion should not be limited to the outcome. An example for applying wrong standards could be setting up a psycholinguistic experiment about language acquisition of children in a kindergarden without briefing and getting the consent of their parents. Doing the research itself may involve hiring helpers, who should not be kept in unethical work conditions; crowdsourcing has been criticized to be a form of slavery, for instance (Fort et al., 2011). Recently, crowdsourcing has become a common element of the NLP toolbox to create gold data or to carry out cost-effective evaluations. Crowdsourcing is now ubiquitous, even indispensable for researchers in HCI, cognitive science, psycholinguistics and NLP. And it poses not just tax compliance issues (who issues tax returns for my workers that I do not know?), but also the fact that the mutual anonymity leads to a loose, non-committal relationship between researchers and crowd workers that stand against pride in the of quality of work output or a moral sense duty of care for workers.

Automation (pertains to R as in Resource and T as in target in RPT). Finally, it could be questioned whether NLP in general is unethical *per se*, based on it being an instance of *automation*: any activity that leads to, or contributes to, the loss of jobs that people use to generate their own existential support: The argument that automation destroys jobs is old (Ford, 2015; Susskind and Susskind, 2015), and traditionally, two counterarguments have been presented. Some claim automation relieves workers from menial jobs so they can pursue more interesting work thereafter. However, many may lack the qualifications or intellect, or may at least perceive stress by that perspective of being forced to taking on more and more challenging jobs. Others even see automation as "freeing humans from duty of work" completely, which would be an ethical pro-argument. In reality, most humans like to have work, and may even need it to give their lives a structure and purpose; certainly many people define who they are by what they do professionally. Therefore, taking their work from them without their consent, means taking their dignity from them. It is also often argued that NLP systems merely aim to make human analysts more productive. It is desirable to do so, and then automation would seem morally legitimate. However, in practice, many customers of applications desire automation as a tool to reduce the workforce, because they are under cost pressure.

5 Best Practices

5.1 Ethics Review Board

In order to establish ethical behavior as a default, installing a *process* likely increases the awareness; it assigns responsibility and improves consistency of procedure and outcome. An ethics review board for companies (as already implemented by universities for use in the context of the approval of experiments with human and animal subjects) included in the discussion of new products, services, or planned research experiments should be considered, very much like it already exists in university environments in an experimental natural science context; in the 21st century, experiments with data about people is a proxy for doing experiments with people, as that data affects their lives. Figure 2 shows our proposal for such a process for a hypothetical company or research institution. It shows a vetting process featuring an Ethics Review Board (ERB), which would operate as follows before and after executing research projects, before and after product development, as well as during deployment of a product or service on an ongoing basis (at regular intervals), the ERB gets to review propositions (the "what") and methods (the "how") and either gives its blessing (approve) or not (veto). Ethics stakeholders participate in research, product/service design & development, operations & customer service. Each of them could report to the Chief Informa-

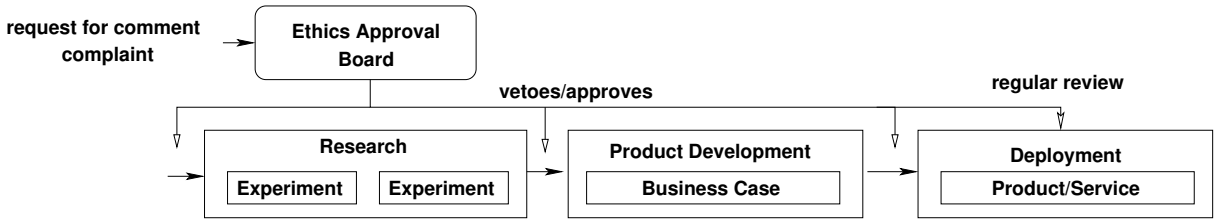

Figure 2: A Process Proposal for "Ethics by Design" in an Organization.

tion Officer via a Vice President for Governance rather than phase-specific management to give the ethics review more robustness through independence. There are associated business benefits (employee identification, reduced risk of reputation damage), however these are distinct from the ethical motive, i.e. desiring to do the right thing for its own sake. They are also distinct from legal motives, because acting legally is often not sufficient for acting ethically, especially in emerging technology areas where the law lags behind technology developments. An ERB might be too expensive to operate for smaller institutions, but it could be outsourced to independent contractors or companies that perform an external audit function ("ERB as a service"). The board would convene at well-defined points to sign off that there was an ethics conversation, documenting identified issues and recommending resolution pathways. ERB audits could benefit from check-lists that are collated in the organization based on the experience obtained in past projects (Smiley et al., 2017). In the US, Institutional Review Boards are already legally required for certain kinds of institutions and businesses (Enfield and Truwit, 2008; Pope, 2009). Our proposal is to adopt similar practices, and to customize them to accommodate particular, NLP-specific issues. An ERB should be empowered to veto new products or NLP projects on ethical grounds at the planning stage or at project launch time (earlier means less money wasted). The ERB could be installed by the board to make the company more sustainable, and more attractive to ethical investors. Note that it is not required that ERB members agree on one and the same school of ethics: a diverse ERB with voting procedures, comprising members, each of which driven by their own conscience, might converge towards wise decisions, and that may be the best way to adopt as a practical solution ("crowd wisdom"). The ethics board should ideally contain all stakeholder groups: if the organization's NLP projects are mostly pertaining to automation as an issue, worker representatives would be good to include. Moral philosophers, human rights experts and lawyers could be included as well; in general, some delegates should be independent and should have a well-developed conscience.[10] In practice, the ERB involvement incorporates elements from "value sensitive design" (Friedman et al., 2008) (thus generalizing Cavoukian's "privacy by design" idea) and works as follows: a product manager generates a document outlining a new project, or a scientist creates an idea for a new research project. At the decision point (launch or nor), the ERB is involved to give its ethics approval (in addition to other, already existing functions, e.g. finance or strategy). At this stage, the morality of the overall idea is assessed. Working at a conceptual level, the ERB needs to identify the stakeholders, both direct and indirect, as well as the values that are implicated in the proposed idea or project. Once the project is launched, detailed written specifications are usually produced. They again are brought to the ERB for review. The project plan itself is reviewed by the ERB with a view to scrutinizing research methods, understanding how the stakeholders prioritize implicated values and trade-offs between competing values, as well as how the technology supports certain values. The ERB may send documents back with additional requests to clarify particular aspects. The ERB documents permanently anything identified as unethical. Ideally, they would have a powerful veto right, but different implementations are thinkable. Much harder is the involvement in ongoing review activities, for example to decide whether or not code is ethical. It appears that a committee meeting is not well suited to ascertain moral principles are adhered to; a better way could be if the ERB was in regular informal touch

[10] It would definitely help to include more inexperienced and younger members, whose idealism may not have been corrupted by too much exposure to so-called "real life".

with scientists and developers in order to probe the team with the right questions. For example, a mention of the use of crowdsourcing could trigger a suggestion to pay the legal minimal hourly salary.

5.2 Responses to Ethics Dilemmas

A lot of the literature focuses on how to decide what is ethical or not. While this is obviously a core question, the discussion must not rest there: of similar relevance is an elaboration about possible remedies. Table 1 shows a set of possible responses to ethics issues. Some of these acts are about the individual's response to a situation with possible ethical implications in order to avoid becoming co- responsible/complicit, whereas others are more outcome-oriented. They are loosely oriented from least (bottom) to most (top) serious, and include internal and external activities.

6 Discussion

The privacy issue in the early UNIX **spell(1)** tool differs from the Mexican propaganda chat-bots in that the former wrongly (un-ethically) implements a particular function (if there is no on-screen warning to ensure informed consent of the user), whereas in the latter case, the chat-bot application as a whole is to be rejected on moral grounds. We can use these two situations to anecdotally test our "ethics by design" process in a thought experiment: what if both situations arose in an organization implementing the process as described above? The spell tool's hidden emails should be unearthed in the "review" stage (which could well include code reviews by independent consultants or developers in cases where questionable practices have been unearthed or suspected). And clearly, the Mexican bots should be rejected by the ERB at the planning stage. By way of self-criticism, flagging functional issues like the hidden spell email feature is perhaps less likely detectable than application-level ethical issues, since over-keen programmers may either forget, intentionally not communicate, or mis-assess the importance of the hidden-email property; nevertheless, using the process arguably makes it more likely to detect both issues than without using it. Floridi's model, which wass designed for information ethics, may have to be extended in the direction of information processing ethics (covering the software that creates or co-creates with humans the information under consideration), since the software or the process that leads to the software can itself be unethical in part or as a whole. There is also an interaction between the conversation whether AI (including NLP) can/should even aspire doing what it does, as it does, because framing of the task brings ethical baggage that some see as distraction from other (more?) important issues: as Lanier (2014) points out, the directional aspiration and framing of AI as a movement that either aims to or accidentally results in replacing humans or superseding the human race, effectively creating a post-human software-only species ("the end of human agency") is a "non-optimal" way of looking at the dangers of AI; in his view, it adds a layer of distractive arguments (e.g. ethical/religious ones about whether we should do this) that divert the discourse from other, more pressing conversations, such as over-promising (and not delivering), which leads to subsequent funding cuts ("AI winter"). We will likely be producing a world that may be likened to Brazil rather than Skynet. While Lanier has a point regarding over-selling, in our view the ethical problems need to be addressed regardless, but his argument helps to order them by immediate urgency. One could argue that our ERB proposal may slow innovation within an organization. However, it is central to protecting the organization from situations that have a significant impact on its reputations and its customers, hence, reducing the organization's risk exposure. One may also argue that, if implemented well, it could guide innovation processes towards ethical innovation.

7 Summary & Conclusion

In this paper, we have discussed some of the ethical issues associated with NLP and related techniques like machine learning. We proposed an "ethics by design" approach and we presented a new process model for ethics reviews in companies or research institutions, similar to ethics review boards that in universities must approve experiments with animal and human subjects.[11] We also presented a list of remedies that researchers can consider when facing ethical dilemmas. In future work, professional codes of conduct should be strengthened, and compliance be made mandatory for professionals.

[11] See also the journal *IRB: Ethics & Human Research* (currently in its 39th volume) dedicated to related topics in other disciplines.

Table 1: Remedies: Pyramid of Possible Responses to Unethical Behavior.

Demonstration	to effect a change in society by public activism
Disclosure	to document/to reveal injustice to regulators, the police, investigative journalists ("Look what they do!", "Stop what they do!")
Resignation	to distance oneself III ("I should not/cannot be part of this.')
Persuasion	to influence in order to halt non-ethical activity ("Our organization should not do this.")
Rejection	to distance oneself II; to deny participation; conscientious objection ("I can't do this.")
Escalation	raise with senior management/ethics boards ("You may not know what is going on here.")
Voicing dissent	to distance oneself I ("This project is wrong.")
Documentation	ensure all the facts, plans and potential and actual issues are preserved.

Acknowledgments. The authors would like to express their gratitude to Aaron J. Mengelkamp, Frank Schilder, Isabelle Moulinier, and Lucas Carstens for pointers and discussions, and to Khalid Al-Kofahi for supporting this work. We would also like to express our gratitude for the detailed constructive feedback from the anonymous reviewers.

References

ACM. 1992. ACM Code of Ethics and Professional Conduct. Online, cited 2016-12-27, http://bit.ly/2kbdhoD.

ACM. 2017. Principles for Algorithmic Transparency and Accountability. online, cited 2017-01-15, http://bit.ly/2jVROTx.

Colin Allen, Iva Smit, and Wendell Wallach. 2005. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7(3):149–155.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4(463–476).

Aristotle. 1925. *The Nicomachean Ethics: Translated with an Introduction.* Oxford University Press, Oxford, England, UK.

Aristotle. 1934. *Aristotle in 23 Volumes*, volume 19. Heinemann, London, England.

Harry Armstrong. 2015. Machines that learn in the wild: Machine learning capabilities, limitations and implications. Technical report, Nesta, London, England.

Association of Internet Researchers. 2017, to appear. *AoIR 2017: Networked Publics – The 18th annual meeting of the Association of Internet Researchers will be held in Tartu, Estonia, 18-21 October 2017.*

Jon Bentley. 1986. *Programming Pearls*. Addison-Wesley, Reading, MA, USA.

Chris Brew. 2016. Classifying ReachOut posts with a radial basis function SVM. In Kristy Hollingshead and Lyle H. Ungar, editors, *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology, CLPsych@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA*, pages 138–142. Association for Computational Linguistics.

Brey, Philip, and Johnny H. Soraker. 2009. Philosophy of computing and information technology. In Dov M. Gabbay, Antonie Neijers, and John Woods, editors, *Philosophy of Technology and Engineering Sciences*, volume 9, pages 1341–1408. North Holland, Burlington, MA, USA.

Terrell Ward Bynum and Simon Rogerson. 2004. *Computer Ethics and Professional Responsibility: Introductory Text and Readings*. Wiley-Blackwell, Malden, MA, USA.

Terell Bynum. 2008. Computer and information ethics. In *Stanford Encyclopedia of Philosophy*. Stanford University.

C. Cary, H. J. Wen, and P. Mahatanankoon. 2003. Data mining: Consumer privacy, ethical policy, and system development practices. *Human Systems Management*, 22.

Ann Cavoukian. 2009. Privacy by design. Technical report, The Office of the Information and Privacy Commissioner of Ontario, Toronto, Ontario, Canada.

Kord Davis and Doug Patterson. 2012. *Ethics of Big Data: Balancing Risk and Innovation*. O'Reilly, Sebastopol, CA, USA.

K. B. Enfield and J. D. Truwit. 2008. The purpose, composition and function of an institutional review board: Balancing priorities. *Respiratory Care*, pages 1330–1336.

Joshua Fairfield and Hannah Shtein. 2014. Big data, big problems: Emerging issues in the ethics of data science and journalism. *Journal of Mass Media Ethics*, 29(1):38–51.

FATML. 2014. Fairness, accountability, and transparency in machine learning (FATML).

Luciano Floridi. 2013. *The Ethics of Information*. Oxford University Press, Oxford, England, UK.

Martin Ford. 2015. *The Rise of the Robots: Technology and the Threat of Mass Unemployment*. Oneworld, New York, NY, USA.

Karn Fort and Alain Couillault. 2016. Yes, we care! results of the ethics and natural language processing surveys. In *10th edition of the Language Resources and Evaluation Conference, 23-28 May 2016, Portoro, Slovenia*, LREC 2016, pages 1593–1600.

K. Fort, G. Adda, and K.B. Cohen. 2011. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Batya Friedman, Peter H. Kahn Jr., and Alan Borning. 2008. Value sensitive design and information systems. In Kenneth Einar Himma and Herman T. Tavani, editors, *The Handbook of Information and Computer Ethics*, chapter 4, pages 69–101. Willey, Hoboken, NJ, USA.

Harry J. Gensler. 2011. *Ethics: A Contemporary Introduction*. Routledge Contemporary Introductions to Philosophy. Routledge, London, 2nd edition.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Immanuel Kant. 1785. *Grundlegung zur Metaphysik der Sitten*. J. F. Hartknoch, Riga, Latvia.

Bryan Klimt and Yiming Yang. 2004. The Enron corpus: A new dataset for email classification research. In J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, editors, *Machine Learning: ECML 2004, 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004, Proceedings*, volume 3201 of *Lecture Notes in Computer Science*, pages 217–226, Heidelberg, Germany. Springer.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, editors. 2014. *Privacy, Big Data and the Public Good*. Cambridge University Press, New York, NY, USA.

Jason Lanier. 2014. The myth of AI: A conversation with Jaron Lanier. Online, cited 2017-01-17, https://www.edge.org/conversation/the-myth-of-ai.

Annette Markham and Elizabeth Buchanan. 2012. Ethical decision-making and Internet research: Recommendations from the AoIR ethics working committee (version 2.0). Technical report, Association of Internet Researchers.

Judith Masthoff. 2011. Review of: Close Engagements with Artificial Companions: Key Social, Psychological, Ethical, and Design Issues. *Computational Linguistics*, 37(2):399–402.

Aaron Mengelkamp, Sebastian Rohmann, and Matthias Schumann. 2016. Credit assessment based on user generated content: State of research. In Christine Bernadas and Delphine Minchella, editors, *Proceedings of the 3rd European Conference on Social Media, 12-13 July 2016*, pages 223–231, Caen, France.

John Stuart Mill. 1879. *Utilitarianism*. Floating Press, Auckland, New Zealand, 1st edition.

Frank Pasquale. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, Cambridge, MA, USA.

Jonathon W. Penney. 2016. Chilling effects: Online surveillance and Wikipedia use. *Berkeley Law Review*, 31(1):117–182.

John P. Pestian, Pawel Matykiewicz, and Michelle Linn-Gust. 2012. Suicide note sentiment classification: A supervised approach augmented by Web data. *Biomedical Informatics Insights*, 5 (Suppl. 1):1–6.

T. M. Pope. 2009. Multi-institutional healthcare ethics committees: The procedurally fair internal dispute resolution mechanism. *Campbell Law Review*, 31:257–331.

Michael J. Quinn. 2016. *Ethics for the Information Age*. Pearson, 7th edition.

Charese Smiley, Frank Schilder, Vassilis Plachouras, and Jochen L. Leidner. 2017. Say the right thing right: Ethics issues in natural language generation systems. In *Proceedings of the Workshop on Ethics & NLP held at the EACL Conference, April 3-7, 2017*, Valencia, Spain. ACL.

Bernd Carsten Stahl, Job Timmmermans, and Brent Daniel Mittelstadt. 2016. The ethics of computing: A survey of the computing-oriented literature. *ACM Computing Survey*, 48(4):55:1–55:38.

Richard Susskind and Daniel Susskind. 2015. *The Future of the Professions: How Technology Will Transform the Work of Human Experts Hardcover*. Oxford University Press, New York, NY, USA.

Andree Thieltges, Florian Schmidt, and Simon Hegelich. 2016. The devils triangle: Ethical considerations on developing bot detection methods. In *Proc. 2016 AAAI Spring Symposium, Stanford University, March 2123, 2016*, pages 253–257. AAAI.

Alan F. Westin. 1967. *Privacy and Freedom*. Atheneum, New York, NY, USA, 1st edition.

Norbert Wiener. 1954. *The Human Use of Human Beings*. Houghton Mifflin, Boston, MA, USA.

Yorick Wilks, editor. 2010. *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical, and Design Issues*. John Benjamins, Amsterdam, The Netherlands.

Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040.

M. Zimmer. 2010. 'but the data is already public': On the ethics of research in Facebook. *Ethics and Information Technology*, 12(4):313–325.