# Ethics and Good Practice in Computational Paralinguistics

Anton Batliner, Simone Hantke, and Björn Schuller, *Fellow, IEEE*

**Abstract**—With the advent of 'heavy Artificial Intelligence' – big data, deep learning, and ubiquitous use of the internet, ethical considerations are widely dealt with in public discussions and governmental bodies. Within Computational Paralinguistics with its manifold topics and possible applications (modelling of long-term, medium-term, and short-term traits and states such as personality, emotion, or speech pathology), we have not yet seen that many contributions. In this article, we try to set the scene by (1) giving a short overview of ethics and privacy, (2) describing the field of Computational Paralinguistics, its history and exemplary use cases, as well as (de-)anonymisation and peculiarities of speech and text data, and (3) proposing rules for good practice in the field, such as choosing the right performance measure, and accounting for representativity and interpretability.

**Index Terms**—Computational Paralinguistics, Good Practice, Ethics, Privacy, (De-)Anonymisation, Representativity, Common Points of Reference, Interpretability, Measure of Goodness.

## 1 Introduction

*"Quidquid agis prudenter agas et respice finem."*
Whatever you do, follow the rules of good practice and consider the consequences for individuals and society.

For decades, ethics and privacy have not been topics within phonetics and linguistics researchers really had to be aware of; this is in strong contrast to other fields such as medical sciences or psychology. Most of the time, there were no institutionalised regulations and committees; more or less 'informed' consent was rather informally given by participants of studies (experimental subjects) and not formally required by universities and funding bodies. All this holds for computer science and engineering as well, when speech and language had been addressed. Thus, the remark by Cowie [1] on affective computing can be extended to phonetics and linguistics, and to speech and language processing in general as well: "People who work in affective computing tend to have trained in disciplines allied to engineering and mathematics. Training in those areas is unlikely to have included courses on ethics. As a result, it can come as a shock to discover that ethical issues are very much part of the discipline that they have come into ..."; cf. [2] as well.

During the last years, ethics and privacy have been playing an increasing role in the scientific and societal discourse, due to both utopic and dystopic visions of new developments – big data, Internet-of-Things, Deep Neural Networks (DNN), availability of personal data on the web, and new possibilities within automatic speech and language processing. This concern is mirrored in discussions, new regulations, and guidelines on all societal levels – EU, states, funding bodies, universities, departments, and newly established ethical committees. Yet, there is no consensus on the good and the bad the new technologies can offer now and especially in the future. There are claims of startups that they are able or will be able in the near future to de-anonymise speakers and to predict – not only diagnose – pathologies such as Parkinson's Disease or Attention Deficit Disorder from yet non-affected speech. This might be as unlikely as the claims that lie detectors really work [3] but of course, it is always difficult to *prove* that black swans or Yetis do not exist, or to argue that the existence of some black swans does not mean that you will encounter them. Consequently, such concerns might find their way into ethical regulations and have to be counterbalanced by concerns that research should not be blocked unduly.

In daily life, researchers inevitably get into contact with ethics when it comes to make a project pass ethical clearance and to get informed consent from participants in experiments. Taking care of such privacy issues is both inconvenient and necessary: inconvenient because it is (at least, has been) a new requirement; necessary because of all the threats and fears that are associated with this topic. Yet, ethics does not only mean privacy but good and responsible research as well – in our case, research in the field of Computational Paralinguistics (CP) where we are interested in the (type of) speech and speaker we are dealing with; cf. Section 3. Note that there is considerable overlap between CP and affective computing; cf. Fig. 2 below.

The focus of this contribution is twofold: First, we want to introduce CP as a field with specific ethical demands – similar to but not identical with the demands in neighbouring fields. Second, we want to point out requirements on good practice in CP: Good research is more ethically defensible than bad research [4], [5], [6]. We concentrate on aspects of speech and language addressed in CP and not on generic topics in ethics and privacy.

To start with, we shortly introduce key theories, con-

- *A. Batliner is with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany, and with the Pattern Recognition Lab, FAU Erlangen-Nürnberg, Germany*
  *E-mail: anton.batliner@lrz.uni-muenchen.de*
- *S. Hantke is with audEERING GmbH, Gilching, Germany*
- *B. Schuller is with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany, with the Group on Language, Audio & Music (GLAM), Department of Computing, Imperial College London, UK, and with audEERING GmbH, Gilching, Germany*

*Manuscript received Month day, year; revised Month day, year.*

cepts, and fields of ethics in Section 2. This is followed by a definition of CP in Section 3.1, by some exemplary use cases in Section 3.2, and by the development of CP in Section 3.3, which has direct impact on ethical demands; Section 3.4 deals with (de-)anonymisation of personal information in speech and text. Section 4 details specific aspects of good practice in CP, such as representativity, common points of reference, interpretability, and adequate performance measures. Section 5 tries to wrap up the different and sometimes contrasting needs of ethically responsible research and applications. Concluding remarks are given in Section 6.[1]

## 2 Ethics

Here, we want to present theories and concepts in ethics that are important in the context of CP, following the structure of Fig. 1. There, groups of ethical theories and concepts are given in capitals and specific theories in boldface, with exemplifications in italics. Specific fields of ethics are framed. Grey background characterises general theories; fields that are put into practice are given with a blue background. Good practice will be introduced in general terms and defined and dealt with in detail in Section 4.

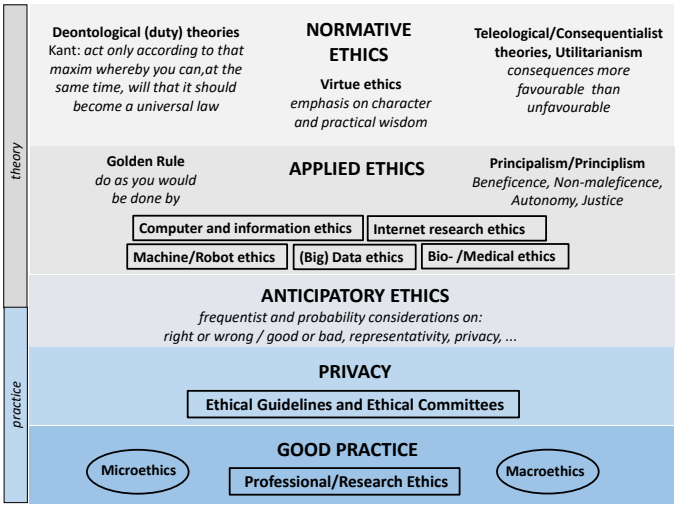


Fig. 1. Ethics: theory and practice

### 2.1 Theories of Ethics

**Ethics** is defined in the Encyclopedia Britannica [9] as follows: "Ethics, also called MORAL PHILOSOPHY, the discipline concerned with what is morally good and bad, right and wrong. The term is also applied to any system or theory of moral values or principles." It is not occupied with factual knowledge but with values, and it is concerned with practical decisions in many disciplines.

**Normative ethics** [10] wants to establish standards of wrong and right behaviours, encoded in one rule or a set of rules. A well-known rule within **Deontological (duty) theories** [11] is Kant's imperative, given in Fig. 1. This is opposed to **Teleological/Consequentialist** theories, **Utilitarianism** perhaps being the most pronounced one: Moral rightness depends only on the consequences (maximising the good and avoiding the bad) [12]. **Virtue ethics** is the third of the major approaches in normative ethics; whereas the other two theories concentrate on rules and consequences, this theory concentrates on virtue and practical wisdom (character, moral) [13]. These normative theories have their own principles but at the same time, they consider characteristics found in the other theories. The same can be said about other ethics theories; for example, Care ethics emphasises "personal interaction and dependency" [14], stresses a feminist perspective, and distinguishes itself from the other normative theories and the concept of justice. Yet, this might be considered rather a matter of attaching importance to different aspects than a fundamental difference.

**Applied ethics** [15], [16], [17], on the one hand, tries to formulate rules and principles that are founded in normative ethics, without necessarily being grounded on and in line with one of the theories; well-known and influential are the **golden rule** (*do as you would be done by*) [1] and the four principles in bioethics [18] that are characterised in [19] as **Principalism** (also called **Principlism** [20]): **Beneficence** "...implies an obligation to do good for your patient." **Non-maleficence** "...implies a duty to do no harm." **Autonomy** "...implies a duty of non-interference, for example, respect for the decision-making capacity of an individual even if the consequences of these decisions are not in their best interests." **Justice** "...is more problematic to define but at its most basic probably concerns access to health care and just distribution of healthcare resources." In [21], justice encompasses fairness, prevention, mitigation of unwanted bias, and discrimination. Autonomy is a central concept: "Individual autonomy is an idea that is generally understood to refer to the capacity to be one's own person, to live one's life according to reasons and motives that are taken as one's own and not the product of manipulative or distorting external forces." [22]. On the other hand, applied ethics branches out into sub-fields – those most relevant for CP are given in Fig. 1: **Computer and information ethics** [23], **Internet research ethics** [24], **Machine/Robot ethics** [25], [26], **(Big) Data ethics** [27], [28], [29], [30], and **Bio-/Medical ethics** [19]. The names of these fields are self-explaining; common to them is, in the context of CP, that the researcher somehow deals with other human beings via the computer, be this by by directly interacting with them, or by collecting and exploiting personal data.

Theories of normative ethics can be self-consistent; this is not possible in applied ethics where authors often resort to common sense [19]. This means as well that, in practice, we cannot do with a fixed set of rules because we cannot foresee exactly what future will bring: **Anticipatory ethics** [31] has to consider the likely outcome; this can be seen as a combination of a frequentist perspective (most likely, most frequent) and a utilitarian perspective (the consequences are good or bad). We will exemplify such anticipatory considerations in Section 3.2.

Ethics is not yet a 'regular' topic within speech research; in practice, however, every researcher gets into contact with ethics when it comes to **Privacy** [32] and by that, to **Ethical Guidelines** that have to be followed and **Ethical Committees** that have to be passed. These topics will be addressed in Sections 2.2 and 2.3. Moreover, practice should be **Good Practice** which is part of **Professional/Research**

1. This article expands the short contribution of [7] and the small Section on 'ethical considerations' in [8].

**ethics**. Practical guidelines for research ethics along the lines of, e. g., principalism are given in [33]. Good practice can mean two different, yet closely connected things: on the one hand, research ethics, on the other hand, simply doing 'good, professional science'. All this can be seen as part of **Microethics** [34] which is **intrinsic** to science [7], [35], whereas general ethical considerations, focusing on individuals, groups, and society at large, can be seen as part of **Macroethics** [34], being **extrinsic** to science[2]. Good practice will be dealt with further below, in Section 4.

### 2.2 Privacy

In the course of history, the possibilities and by that, the importance of privacy developed. This was due to, on the one hand, the spreading of wealth and civilisation – houses in old times had only one room, and the whole family, together with guests, often slept in one big bed; cf. the process of 'Intimisierung und Privatisierung des Schlafens' (Intimisation and privatisation of sleeping) from the 16th to the 19th century [36]. On the other hand, it was a matter of culture; in ancient Rome, there were public rest rooms. The privacy as we know it has not been a long established achievement; it is now threatened by the overall availability of personal data (in the cloud) and the willingness to give them away. (Further information can be found in [37] and [38].) The classic account of privacy is given in [39] where the authors stress the 'right to be left alone'. Their introductory remark is today as relevant as it was in their days: "That the individual shall have full protection in person and in property is a principle as old as the common law; but it has been found necessary from time to time to define anew the exact nature and extent of such protection." This is exactly the situation we are facing today: Normal is the need to balance the interests of the public against the interests of celebrities (royals, film stars, politicians); new is the need to balance the interest of private persons (i. e., of everybody) against the interests of companies and society. In both cases, individuals and groups have to be protected against any data abuse.

As far as participants in experiments (experimental subjects) and, in a broader sense, human informants are concerned, most important is to preserve their privacy to avoid any possible harm by giving away their speech, together with their identity. We will elaborate on the possibilities of (de-)anonymising speech and text data in Section 3.4; there, we will argue that for speech data, the risk is lower than in the case of video data – and this should be taken into account when establishing ethical regulations. Less important in CP is the danger of direct physical or emotional/mental harm as it can be found in clinical studies: A new drug can cause immediate harm – speech recordings normally do not. Yet, this could change if we, e. g., employ intrusive methods for eliciting moods or emotions.

### 2.3 Ethical Guidelines and Ethical Committees

Scientific bodies, e. g., The American Psychological Association (APA) [40] or the Association for Computing Machinery (ACM) [41], as well as political institutions, cf. the Charter of Fundamental Rights of the European Union [42], address the right to privacy and the importance of informed consent.[3] The EU's General Data Protection Regulation (GDPR) "on the protection of natural persons with regard to the processing of personal data" [43], [44] has binding power for the community starting May 25th, 2018. In [21], Jobin et al. find a convergence around ethical principles in Artificial Intelligence (AI) ethics guidelines, however, with "substantive divergence" as far as interpretation and implementation are concerned. A critical evaluation of AI ethics guidelines can be found in [45].

Stacey and Stacey [46] list the following eight principles that are "... common to contemporary research ethics protocols and standards:" (1) informed consent, which implies the avoidance of covert or secret participant observation; (2) privacy of participants (confidentiality and anonymity); (3) avoiding harm (including psychological effect) and doing good; (4) cognisance of vulnerable groups; (5) participants' right to withdraw or terminate; (6) restricted use of data; (7) due care in the storage of data; (8) avoidance of conflicts of interest. This is a fairly complete catalogue of principles to be followed; yet, importance differs across disciplines: As mentioned above, immediate harm is less likely in CP. Further information on ethical guidelines for related fields can be found in [47] (counselling and psychotherapy), [48] (research with children), and in [49], [50], [51] (speech therapy).

Ethical awareness is a moving target and rules of conduct are being redefined and refined regularly. Its importance differs between disciplines, evolved during time, and has not yet been fully constituted so far: It can be that universities without medical departments have not yet established an Ethics/Ethical Committee (EC), and/or ECs are still discussing their rules of conduct and adapting them – more or less in an ad hoc manner – to proposals they have to evaluate. In other scientific fields, there has been a longer tradition of ECs and a critical discussion of implications as well. For instance, for sociological and ethnographic research, Schrag [52] argues against the 'ethical imperialism' in Institutional Review Boards (IRB)[4]: "... compared to the problems of medical research, serious social-science abuses are quite rare". This is corroborated in [53]: "... although the possibility of harm to participants in ethnographic research is real, the probability of harm is very low". Alternatives to the traditional IRBs are discussed in [54] and in [55]: "Chief among these points is the importance and right to conduct research as a vital element for a democratic society that values the freedom of expression ...". Research is, of course, not foremost the right of a few researchers but its results can be important and beneficial for society – and all individuals that are part of this society – as well.

The question of how to handle older databases without a full-fledged ethical clearance seems to be largely unexplored. When will privacy requirements expire and the

2. " 'Microethics' considers individuals and internal relations of the engineering profession; 'macroethics' applies to the collective social responsibility of the profession and to societal decisions about technology." [34]

3. The URL: http://ethicscodescollection.org/, retrieved 05/07/2020, provides "the largest online repository of ethics codes and guidelines in the world".

4. The term EC is used more often in the EU, whereas the term IRB is more common in the US. Other terms exist as well, e. g., ethical review board (ERB) or research ethics board (REB).

data be part of history (sort of copyright agreements), and most important, which types of data require which types of privacy rules? For clinical studies, there are established rules for different types of experimental treatments that might possibly be more harmful than other types. It seems that similar distinctions are just about to emerge in humanities and social sciences, and especially in the field of CP; see below Fig. 5.

## 3 Computational Paralinguistics

### 3.1 Definition

The first hurdle we have to face when defining a scientific field is that definitions are notoriously fuzzy when it comes to the border regions separating different fields that might, at first sight, be unambiguous if we only look at the prototypical core area. For us, the problem starts with telling apart speech from language: partly same or different? We will use definitions based on the scientific sub-cultures that have evolved in the course of the last fifty years: *Automatic Speech Processing (ASP)* deals with *spoken language*, *Natural Language Processing (NLP)* deals with *written language* but extends to spoken language when its linguistics is dealt with. Both address the question *what* has been produced, i.e., what has been spoken or written: phones (underlying: phonemes), words and sequences of words (n-grams, collocations), or the semantics behind these words such as keywords, topic spotting, or ontologies. *How* something has been spoken – in which tone of voice, using which prosody, employing specific words out of several candidates that denote the same but have different connotations – all this we attribute to the field of *Computational Paralinguistics (CP)*; 'Automatic' in ASP and 'Computational' in CP both simply mean that the job is done with the help of or by the computer.

Note that extensionally, all these fields have been defined differently in different sub-cultures as well: Sometimes, speech processing is seen as a sub-field of language processing; sometimes, paralinguistics is confined to non-verbal aspects of speech, leaving aside verbal/linguistic aspects. Our motivation to use a rather broad definition of CP is not to annex as much as possible under this heading, but simply to mirror daily practice within the computational approaches towards paralinguistics, or, in other words, to base our definition on *functional* (what do we want to find out) and not *formal* (which means do we use) aspects [8].

Fig. 2 displays the three vocal/verbal aspects that constitute CP in quadrants I, II, and III. Traditionally, paralinguistics encompasses often more than vocal and/or verbal aspects, cf. the fourth quadrant in Fig. 2. Yet, we follow the definition in [8] and constrain paralinguistics to verbal and/or vocal aspects. Thereby, the extensional definition is unequivocal: Everything belongs to CP that has been produced by a human[5] and can be recorded with a microphone and/or represented in a written text. The main topic of CP is not denotation (what we are talking about) but connotation (how we are taking about it). The first quadrant in Fig. 2 *nonvocal & verbal* represents natural

5. Here, we leave aside productions by machines such as avatars or robots that might get a paralinguistic 'flavour' when used in human-machine-communications.

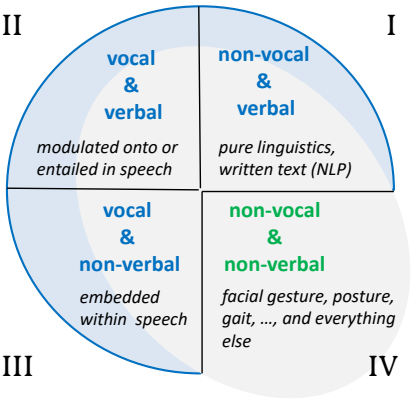


Fig. 2. Vocal and verbal aspects of Computational Paralinguistics (blue); the area of affective computing is indicated by overlaid grey shadowing

language that traditionally is the field of NLP; within CP, this is the field of *opinion mining* or *sentiment analysis* and of any other area when only written language is analysed. The second and the third quadrant represent the traditional fields of paralinguistics, namely vocal aspects that are either modulated onto or embedded within speech. *Vocal & non-verbal* (quadrant III) events such as laughter, filled pauses, grunts or affect bursts such as *grrr, ihh, umpf* are embedded within the speech chain or isolated; they can be modelled in a similar way as words. In quadrant II, we find *vocal & verbal*: for instance, voice quality or intonation modulated onto or entailed in speech. The choice of different words for the same denotation with different valence (e. g., *lady*, *woman*, or *slut*, all denoting [+human], [+adult], [+female]) belongs to *verbal*, thus, to both quadrants I and II. Automatic speaker identification/verification belongs mostly to *vocal*, thus to both quadrants II and III. CP spans across both Automatic Speech Recognition (ASR)/ASP and NLP, concentrating on psychological, sociological, and medical aspects: What characterises the individual, what is typical for groups and social classes, what tells apart typical from atypical behaviour?

*Non-vocal & non-verbal* in quadrant IV stands for all other modalities such as facial and body gestures, gait, or physiosignals – and in a broader sense, for every other type of context. On the one hand, the context modelled by metadata of corpora – age, gender, social class, and alike – belongs to quadrant IV; on the other hand, of course it has to be taken into account when discussing ethics in CP. Basic considerations will hold across all modalities; technical solutions, however, might differ. The area of *affective computing* is indicated by grey shadowing in Fig. 2; there, we have to consider ethics both when collecting data and recording participants in experiments, and when generating (embodied) agents or robots that interact with humans not only by exchanging information but by exchanging emotions as well. Ethical issues within affective computing are discussed in [1], [2], [19], [56], [57], [58], [59].

### 3.2 Exemplary Use Cases

In this section, we want to present a few exemplary use cases in CP and sketch relevant ethical considerations. Fig. 3

attributes these use cases to a layered taxonomy of CP [8]; biological trait primitives are modified by cultural trait primitives; they all manifest themselves in personality traits and in short-term emotions as well as in medium-term or long-term atypicalities or are modified by them. We distinguish between use cases with primary impact on individuals from those with primary impact on groups. This is evident in the case of speaker identification (one individual) vs screening (a population); yet, there is a smooth transition in the other cases (e. g., subjects with a specific regional accent can be classified and stigmatised as group or as individual).

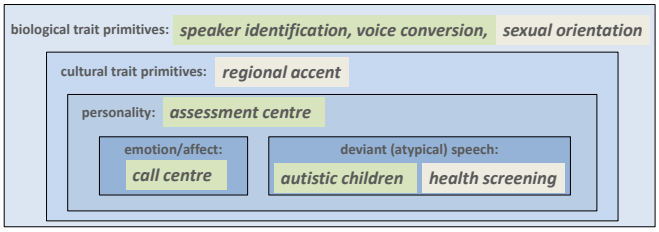


Fig. 3. Taxonomy of Computational Paralinguistics with exemplary use cases; adapted from [8] in a simplified version; green: primary impact on individuals, grey: primary impact on groups

**Speaker identification, forensics/court:** Speaker identification, e. g., in court, can be employed the same way as other evidence; a *voice print* is, however, not at the same level of evidence as finger prints or DNA. This might change if the search space is restricted (the same way as for speaker verification out of a smaller number of candidates); cf. below section 3.4. Privacy can be violated, e. g., with secret recordings at home in the case of divorce proceedings, and speaker identification can be used, in a dictatorial regime, for the prosecution of oppositionists.

**Voice conversion/forgery:** Brundage et al. [60] deal with "the malicious use of artificial intelligence". The authors do not discuss dystopic visions that are projected into the far future but developments that are likely to emerge in the next five years or have emerged already. Amongst them is *visual forgery* ("... the ability to generate synthetic images, text, and audio could be used to impersonate others online, ...") or *spear phishing* ("... the attacker often posing as one of the target's friends, colleagues, or professional contacts."). *Voice conversion* [61] and by that, *voice forgery* seem to be on the verge of being technically possible as well and available to the public. At present, it should still be possible to detect such a manipulation. This might change in the future and might lead to some disruption point where both privacy and anonymity are challenged and at the same time, every speaker can be imitated close to perfection; it might still be possible to detect such a voice forgery but it can of course be very harmful even if its validity can be questioned.

**Sexual orientation:** This is a highly private issue. Even if equal rights should have been established, this is not fully the case in Western societies; moreover, diverse people (LGBT: Lesbian, Gay, Bisexual, Transgender) are still prosecuted in some societies. Kosinski et al. and Wang et al. [62], [63] claim that private traits, including sexual orientation, can be detected automatically via big data (Facebook, digitally available images, etc.) with a performance of over 80% correct.[6] We do not know of any attempt to do this employing acoustic information. Yet, studies such as [66], [67], [68], [69] demonstrate that this could be possible as well.

**Regional accent:** This is, in itself, a highly interesting research topic [70] but its recognition can have more serious consequences than a wrong target pronunciation in language classes: The prediction of neighbourhood someone is living in can be used to restrict access to loan, or to decide upon probabilities that prisoners will re-offend [71].

**Assessment centre:** Human resources (HR) in big companies already use voice and linguistic cues for personality testing. At best, this might do no more harm than personal likes or dislikes of HR officers if used as one amongst other criteria. Yet, it should be clear that the performance claimed by companies marketing such products is more than doubtful and not backed up by scientific evidence [72], [73], [74].

**Call centre:** Estimating the length of conversation as a measure of felicity/success, or estimating the interest of the caller with the help of audio features, is ethically not critical. However, trying to detect callers' anger or personality, or trying to assess agents' friendliness might be OK, if anonymous, but highly critical, if this information is harnessed for decisions on employment.

**Autistic children:** In the ASC project [75], the children had to look at and listen to actors producing different emotions, and then, they had to produce these emotions themselves. Here, what is the criterion for right or wrong? A real hit, that is, a correspondence with the prototypical acted emotion? Or the produced emotion can be perceived as such but is not very pronounced/typical? Or it is rather atypical and ambiguous but not outright wrong, indicating another, opposite emotion (happy instead of angry)? And most important, in the case of erroneous recognition or classification, when we teach awkward or wrong expressions of emotions, risks are high that the outcome of such a therapeutic game is not only irrelevant but outright harmful, leading to social disintegration. This is an example for applications that are aimed at individual patients (here: children) with possible ethically critical impact for these individuals [48], [76].

**Health screening:** Automatic screening of health state via voice [77] can be a valuable instrument in the case of speech therapy; however, we could imagine that health insurances or employers use it for decisions whether to admit or employ candidates. This can be unethical, even if the diagnosis is correct, and it can be disastrous if the diagnosis is wrong.

Table 1 illustrates that methods themselves are neutral: They can be harnessed for doing good things or bad things. Note that 'bad' can mean 'bad if false alarms' and/or 'bad if hits'. The vague specification 'mostly' good or bad in Table 1 relates to the fact that, without further specification, we cannot say that some use is always good or always bad. When we look at the call centre use case: It might get

6. This study has been heavily criticised [64] for confusing biological indications with social markers. In fact, the problem is not necessarily that social proxies (i. e., substitutes, cf. Section 4.2) are employed for classification but that it is done at all, that possible stigmatisation can affect both correctly and wrongly classified people, and that physiognomy is taken up again [65].

TABLE 1
Use Cases: The Good and the Bad

| USE-CASE | (mostly) GOOD | (mostly) BAD |
|---|---|---|
| speaker identification, forensics/court | criminals (blackmailer, terrorist) | divorce proceedings, oppositionist |
| sexual orientation | research | de-anonymisation, stigmatisation |
| voice conversion | entertainment | fake news, etc. |
| regional accent | pure research | acceptance of credit |
| assessment centre | personality scores | personality scores |
| call centre | costumer gets angry, action is taken | agent is unfriendly and gets dismissed |
| autistic children | social integration | social disintegration |
| health screening | early detection of diseases | used by insurances, other illegal exploitation |

rather common to monitor a user's emotional state in this scenario. This can lead to better communication. However, some users might simply not like this and might object to such a monitoring. Moreover, harnessing personal traits (such as emotional behaviour) together with other personal information might be highly unfavourable for the user. In the same vein, monitoring the call centre agents' emotions can be very helpful for them if this is done in the course of a training phase but can have bad consequences for them if it is used to eventually dismiss specific agents because they 'cannot control their emotions'. If this is based on a correct classification, it might violate the agent's privacy; if this is based on a wrong classification (false alarm), then it is wrong per se. Thus, all specifications about 'good' or 'bad' are based on some basic reasoning (similar to but not identical with normative rules), followed by some anticipatory reasoning based on past frequencies (call centre companies behaved mostly in this or that specific way) yielding probabilities for future consequences (utilitarian perspective). Similar constellations can be found for the other use cases we describe in this section. We will come back to use cases and applications in Section 5.2.

It seems not to be possible to attribute specific use cases to certain principles that have to be taken care of whereas other principles are irrelevant. It is rather more or less weight that has to be given to specific principles for specific scenarios. All applications claim to be beneficent either for a single user or for society. A benefit for society can imply violating the interests of the individual to some extent, for instance, when health screening is performed against the will of the individual. Most important is to assess the possibilities of maleficence which as well can be seen as cover term for other principles such as justice: when justice is violated, we can subsume this under maleficence. Ethics of care, for instance, is most relevant for vulnerable groups such as (atypical) children and minority groups. When individuals are targeted, harm for individuals is in the fore, and vice versa, when groups are targeted, groups are in the fore but of course, the individual belonging to such a group is affected as well. In [78], taxonomies of applications in speech emotion processing, especially taking into account ethical awareness, are given.

### 3.3 Development

ASP looks back at more than fifty years of research [79], [80], starting with the processing of single digits, produced by single speakers, in the 1960ies; the lexicon grew from some 1000 entries in the 70ies to several 1000 in the 80ies; trained dictation in the 90ies was followed by robust processing of millions of words in the first decade of this century; the state of the art approaches real-life recognition and language identification with subsequent automatic translation. NLP evolved on a similar timeline [81], [82].[7] All this holds for the processing of what has been said: the chain of words (i. e., word recognition) and the semantics behind (keywords, topic spotting, hot spots, or ontologies). Within humanities, what has been said is normally dealt with within phonetics and linguistics. Now, we address how something has been said by whom: The term *paralinguistics* dates back to the 50ies [8], the field of CP can be traced back to the recognition/verification/identification of speakers, starting in the 70ies; automatic emotion recognition by using speech emerged in the 90ies and was subsequently complemented by classifying/detecting a plethora of long-term speaker traits (age, height, personality, non-nativeness, dialect, pathology, etc.), of intermediate traits/states (intoxication, sleepiness, etc.), and of short-term states (besides clear emotions: interest, boredom, even heart rate or eye contact by using acoustic information, and alike).

We will now sketch the developments within (ASP/NLP and) CP that led to higher ethical demands:

**From *what* to *how*:** Pure ASP or NLP are interested in *what* has been spoken or written; paralinguistics is interested in *how* speech or written language have been produced: in which emotion, by whom (by a non-native, sleepy, intoxicated, nervous, happy, ... person). It is evident that the extension from *what* to *how* something has been produced opens new challenges for ethically acceptable approaches.

**From basic research to application:** Pure speech research in contrast to clinical studies might be considered to do no harm as long as the privacy of (experimental) subjects is guaranteed. Of course, things change if it comes to using results in political debates or to decisions that have direct or indirect impact on sub-populations or individuals such as acceptance or rejection of specific therapies. Applications, on the other hand, if they are not only entertainment or harmless edutainment (in the sense of [57] 'ethically lightweight'), can have serious impact on individuals.

**From typical to atypical:** Typicality is a fuzzy and layered concept [8]: It can mean 'prototypicality' in the sense of 'extreme, very pronounced', thus rather infrequent; it can mean 'very frequent' in the sense of 'typical for a specific (sub-)population' and thus (mostly) less pronounced. Moreover, it can mean both. In the context of speech processing, 'typical' often simply means that data are easily obtainable – in contrast to 'atypical' ones. Often, a 'typical'

7. Note, however, that this performance is still far from the competence of a native speaker.

characteristic is not very interesting for paralinguistics; it is rather the deviation, the *atypicality*, which is interesting. Yet, we always need typical, neutral data (as a sort of background model) in order to find out what is deviant, atypical. For ASR, children are atypical because there is still way more recorded speech data from adult speakers. Within the groups of children, there are again typical children and atypical ones, for instance, those with autism condition or attention deficit disorder, and within these sub-groups, there are again representative, typical children and those at the edges of the distribution. Building smaller sub-samples consisting of people with atypical characteristics is a prestage of personalisation: It is easier to find an individual out of a small group as opposed to finding an individual out of a million people. Ethical considerations already start with the names given to these atypical groups to ensure politically acceptable ('correct') terms, cf. 'autism' vs 'autistic spectrum' ('autism disorder' vs 'autism condition') or the development from 'coloured' to 'negro' to 'black' to 'African American' [83] to 'people of colour'.

**From recognition to analysis:** CP systems so far are mainly tailored to do the recognition job; this means that they only target the assignment of a label to a speech or text unit. A very recent and likely future trend, however, is to go beyond and provide additional analysis, such as the confidence level or prototypicality, regulation, feigning, display rules, or atypicality. This can go as far as to the feature level by analysing which (acoustic or linguistic) feature is different from the standard case in which way. Such research strategies might contribute to de-anonymisation as well.

**From uni-modal to multi-modal processing:** In our definition of paralinguistics, it is confined to verbal/vocal (non-verbal) and written phenomena; in a broader sense, it encompasses other modalities as well, such as facial expressions, hand/body gestures, and gait. Notwithstanding this definition, ethical considerations become more important when multi-modality comes into the game, simply because personalisation is easier; thus, anonymisation has to be stricter: Video processing is more critical as far as identification of individuals is concerned. Thus, especially when CP is embedded in multi-modal approaches, ethical demands are higher.

**From small data to big data:** Prototypical for a small data study is an experiment with some 10–50 subjects who are recruited from the student population, earning credits for participation, or even from a circle of friends; the participants are known beforehand or registered for the study. Prototypical for a big data study is a very large sample – from a few hundred to several thousand and more subjects – obtained from some external source, e. g., from the web (Facebook, YouTube). These subjects can be known, even well-known if they are celebrities or, e. g., taking part in TV discussions, or they can be unknown. Along the same lines, we can talk of in vitro, lab(-oratory) studies on the one hand, and in vivo, real-life, 'in-the-wild' studies on the other hand. Typically, the former employ a much smaller number of subjects than the latter. For small data, most of the time, anonymisation will be necessary; for big data, most of the time, de-anonymisation has to be prevented. For both small and big data, meta-data and meta-information (biological trait primitives such as age and cultural trait primitives

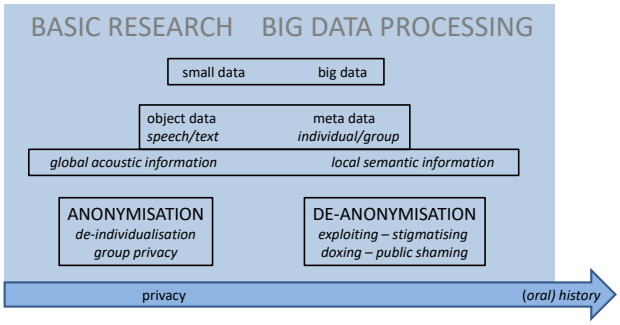


Fig. 4. Anonymisation and De-Anonymisation of Speech Data

such as ethnic/regional background or social class) can be obtained. Both anonymisation and the prevention of de-anonymisation are, as far as privacy considerations are concerned, the most important topics to be dealt with; cf. Section 3.4. Note that CP mostly still deals with small data because the phenomena addressed are inherently vague and have often to be manually annotated by experts in order to get reliable reference classes, in spite of a plethora of possible alternatives such as crowd sourcing, transfer learning, or unsupervised learning.

### 3.4 Anonymisation and De-anonymisation of Speech and Text Data

Fig. 4 summarises the topics of this section that are relevant for both basic research and big data processing, and for the processing of speech and text in CP: types of data – small/big, object/meta, and types of information – global or local. All this is basis and material for the antagonistic attempts towards anonymisation and de-anonymisation. A specific aspect is the time line that might make privacy considerations obsolete when the data are getting part of (oral) history.

In laboratory experiments employing *small data*, the individual participant is known, his/her identity, however, has to be concealed in the following processing, especially when data are passed on to third parties or are made publicly available. Such *basic research*, targeted towards specific phenomena, does not necessarily need information on subjects apart from a broad description of the sample such as age range, language proficiency, or characteristics of pathological traits; else, it is enough when data can be identified unambiguously as belonging to one 'item' or 'subject' in processing. Yet, from a broader perspective, it is advisable to collect as much individual information as possible – data are precious and with their help, it might be possible to address other questions later on. Of course, this conflicts with early anonymisation.

In *big data processing*, the identity of the individual is mostly not immediately apparent but can be discovered. In NLP, the situation was slightly different but the result was the same: It " ... used to involve mostly anonymous corpora, with the goal of enriching linguistic analysis, and was therefore unlikely to raise ethical concerns" [84]. The very same approach of using big anonymous corpora started to raise these concerns when it was possible to de-anonymise single persons. Big data can be anonymous or personalised:

In sentiment analysis, when we 'only' are interested whether some specific product or film receives positive or negative reviews, we do not need any personalised information about the people behind these reviews. Yet, the temptation to find out more is high (see the personalised advertisement in web browsers). In a scan for specific people using big data procedures, e. g., by national security agencies, personalisation is a sine qua non, because the disclosure of individual information provides possibilities to trace back individuals. When online platforms claim that they only maintain anonymous records, this does not mean that a specific person or his browser, computer, network equipment, IP address, or phone cannot be identified: By that, the platforms can associate observed behaviours with the record assigned to individual users and tailor their content and services accordingly. Thus, "...the oxymoronic notion of an anonymous identifier [is] more accurately labelled a pseudonym. These identifiers are anonymous only insofar as they do not depend on traditional categories of identity while still serving the function of persistent identification." [85].

*De-individualisation* (i. e., removing elements that allow data to be connected to one specific person) is just one aspect of *anonymisation*. Location, gender, age, and other information relevant for group membership and thus valuable for statistical analysis relate to the issue of *group privacy*. Thus, anonymisation of data is a matter of degree of how many and which group attributes remain in the data set. To strip data from all elements indicating group membership would mean to strip them from their content. In consequence, despite of the fact that data are anonymous in the sense of being de-individualised, groups – and by that, minority (atypical) groups that are often stigmatised – are always more transparent [30].

Narayanan and Shmatikov [86] report successful *de-anonymisation* attacks "against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on", applied to the Netflix Prize dataset. They challenge in [87] the belief held by "... today's practical practitioners ... [that] records containing sensitive individual data can be 'de-identified' by removing or modifying PII [personally identifiable information]."[8] All this can lead to *doxing*, "... the intentional public release onto the Internet of personal information about an individual by a third party, often with the intent to humiliate, threaten, intimidate, or punish the identified individual ..." with the three types: "de-anonymization, targeting, and delegitimization." [90][9].

Although the data are public, no one really imagines to be the subject of research in Twitter or Facebook studies. Yet, data are collected from social media without considering that the lack of informed consent would in any other form of research constitute a major breach of research ethics. It seems that such a procedure is tolerated as long as the data are not made publicly available – even if they are, e. g., from YouTube. Note that here, we speak about present-day data. It is not clear how old speech/text data have to be to constitute a sort of *oral history* (or written history) data that can be harnessed without any explicit consent from the speaker/writer, i. e., if we simply can and have to treat them along copyright regulations. Thus, the *time frame* we reasonably can overlook is an important factor: We have stressed in Section 1 that the claim that machine learning procedures can predict speech pathologies for not yet affected, i. e., typical speech is not warranted. Young children are a highly protected group with strong requirements to privacy; after 20 years, however, it will be virtually impossible to induce the identity of an adult only from the speech produced by him/her in childhood. We cannot foresee the next centuries to come. Yet, when we are certain that – if at all – this will only be possible after several decades, we might reasonably assume that the speech data can be treated along the lines of oral history by then.

When researchers really want to be on the safe side, they have to stop collecting and publishing data. This will not happen, and it does not make much sense – it would stop serious science and leave the field to less serious players. So it means to assess the pros and the cons in each case. A pivotal aspect will be to weigh costs and benefits: How expensive is it to de-anonymise, and what can I get out of it? Thus it is a matter of balancing probabilities: Which risks do we accept for which benefit [91], [92], [93], [94], [95]?

Now, what about the primary *object data* within CP? Can we anonymise speech? And can we de-anonymise speakers just by analysing their speech – its acoustics and/or its linguistic content, the latter – if orthographically transcribed either by hand or by ASR – being more or less the same as any written record?

The first strategy for anonymising speech databases aims at *globally* manipulating *acoustic information*: Available might not be the raw speech file but extracted features (not low level features such as frame-based MFCCs or pitch but functionals or structured features such as pitch maximum. Glackin et al. [96] propose symbolic encoding with an acoustic model on the client's side, then, the data are sent encrypted to the server. Lopez et al. [97] claim that de-identification with frequency warping and amplitude scaling for depressed speech yields "promising de-identification results at the expense of a slight degradation of depression detection". Encryption for privacy-preserving paralinguistic mining is presented in [98], combined with Support Vector Machines, for emotion recognition, and in [99], combined with neural networks, for health-related tasks. At least DNNs can be fooled by so called 'adversarial examples' (adversarial training) [100], [101], [102] which have been generated from the speech file by just adding some noise or altering some samples. The resulting file cannot be distinguished from the original by humans but creates serious problems for machine learning (ML). By that, ML approaches might not be able to de-anonymise based on speech information. Federate learning [103] distributes training data over a large number of sites, aiming at a 'high-quality centralized model' [104] without making raw data

8. "'personal data' means any information relating to an identified or identifiable natural person (data subject); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;" [88]; cf. as well [89].

9. "The term 'doxing' comes from the phrase 'dropping documents' or 'dropping dox' on someone, which was a form of revenge in 1990s outlaw hacker culture that involved uncovering and revealing the identity of people who fostered anonymity ..." [90]

available; yet, it is vulnerable to poisoning attacks [105]. All these approaches might be suited for specific applications but not for serious research that simply needs the raw data.

The second strategy for anonymising databases aims at *locally* deleting/masking (personal) *semantic information*: In text data, extracted terms are replaced with dummy variables (sanitisation) [106], [107], and in speech data, words or phrases are masked by beep or noise [108]. By that, 'meta-information' such as names or professions that normally are given in the meta-data is concealed. Yet, there is still 'personal information' denoting the individual speaking or writing style.

In [109], there is an in-depth discussion of privacy-preserving technologies in speaker and speech characterisation. An overview of paralinguistic phenomena that can be employed to obtain personal information in speech, and of pertinent literature, is given in [110].

Speech is per se more anonymous than video and at the same time, less attractive for doxing – yet attractive as well for forgery, e. g., in the case of celebrities. Especially threatened by *public shaming* (making private recordings public) are speakers that are not anonymous, such as celebrities (Royal Highnesses, movie stars); it can be doubted that this is attractive in the case of atypical speakers displaying specific traits, for instance, because of (speech) pathologies. Public shaming commonly employs videos – we do not know of any case where speech records were used for this purpose, especially when these speech records were part of scientific databases and shared with other sites; obviously, this is a restricted scenario with too many steps to go for doxing.

Nowadays, speech is believed to be non-anonymous in a strict sense, not only because it entails personal information, but because it creates a *voice print*, the same way as a finger print. This is a utopic belief. Of course, we can narrow down alternatives and/or recognise correctly, especially if meta-data are available. There is another claim that it can be detected whether 'normal', typical speakers will develop some pathologies in the future. So it will be 'seemingly typical but in fact (in future) atypical' speakers; we doubt that the machine is able to detect what an expert cannot, but of course, ML can get better in detecting the transitional area from typical to pathological. And there is no doubt that the machine can accomplish time-consuming tasks such as recordings of babies all night long and for a longer time, including processing of these recordings, in order to detect pathologies at an early stage as in the case of the Rett-syndrome, cf. [111], [112].

Predictions oscillate between realism and utopia, for speech [110] the same way as for other fields. Yet, we can resort to empirical facts: if we again reconsider doxing, to its frequencies based on video or speech or other (meta-) data. And it is getting more concrete when we consider differences between types of information entailed in speech data: Fig. 5 gives an overview of different types of speech data which display more or less personal information. Typical adult speakers are represented often in speech databases; their speech can be recognised best. All other types of speakers are atypical and not represented that often in speech databases; their speech cannot be recognised as easily. However, they might display speech phenomena that can be attributed to smaller groups of speakers (dialects, sociolects, age groups, speech impediments). Given that meta-data are strictly kept confidential (especially the name of the speaker, of course), recordings of read speech and a typical adult speaker are 'most anonymous'; on the other end, recordings of spontaneous speech where the speakers talk about themselves and moreover can be attributed to some specific sub-groups can be considered to be 'less anonymous'. Thus, the likeliness that speech data can be de-anonymised increases from upper left to lower right in Fig. 5; basically, it is lower for speech data than for other types of data, especially image and video data.

| | LESS | MEDIUM | MORE |
|---|---|---|---|
| speech | read | non-prompted but structured | spontaneous, possibly revealing personal information |
| speaker | typical adult | typical sub-group: child, minority, dialect, ... | atypical 'sub- (sub-) group', e. g., pathological speech |
| modality | - | audio only | plus video (face, body, bio-signal) |
| meta-information | only pronunciation | linguistics: words, syntax, semantics | health state, age, weight, height,...; personal information from the web |

Fig. 5. Typology of Speech Data: From less (upper left) to more (lower right) personal information

## 4 Good Practice in Computational Paralinguistics

Good practice in research is a fuzzy concept and used differently in different scientific fields. It can simply mean to follow meaningful rules, to document every decision, and to provide a full account of the data. It can mean to do ethically responsible research such as taking care of privacy and not stigmatising minority groups. A good albeit equally fuzzy definition is: Good practice constitutes good science. Good science is ethically more acceptable than bad science; yet, it is more than 'pure' ethics. Good practice does not wast unduly – mostly public – money and it characterises solid and 'durable' research: When, after a generation, an article is still a good read, then it most likely is good science and the authors followed the rules of good practice.

Fig. 6 displays a sort of flowchart for good practice in CP: A necessary prerequisite is an honest attitude such as no cheating and no plagiarism; this is shortly mentioned in Section 4.1. To start with, we have to collect the 'right', representative data (Section 4.2). Isolated experiments and results can be spurious; thus, we need common points of reference to evaluate our methods (Section 4.3). Measures and actions differ, depending on our aim: For example, do we want to detect or treat an individual, or do we want to screen a population? This is addressed in Section 4.4. The possibility of interpreting our results is pivotal: If we cannot do that, we simply do not know why we decide (Section 4.5). And when we communicate our results to colleagues and to the public, we have to choose metrics that are correct and can be understood at the same time (Section 4.6). Note that aiming at the best possible performance could be seen

as part of good practice as well – it is the Holy Grail in ML. Yet, a lower performance can be ethically more adequate, if results can be interpreted or stigmatisation avoided.

Taking into account all these rules constitutes good practice; yet, in addition, we have to consider both the interests of individuals, groups, and society. We try to depict these aspects in Fig. 6 by including the four principles of Principalism, attributing Autonomy to Privacy, and the other three Beneficence, Non-Maleficence, and Justice, to the need of balancing. Privacy and Balancing are not meant to be two independent 'ethical modules'[10]: They are highly intertwined but told apart. In practice, individual privacy has to be taken care of simply because research has to pass ethical clearance. Balancing relates not only to the individual (for instance, in the role of experimental subject) but to groups and society at large as well. Note that privacy preserving research is not necessarily good research in the sense of good practice that can be seen as intrinsic to science, with impact on ethics.

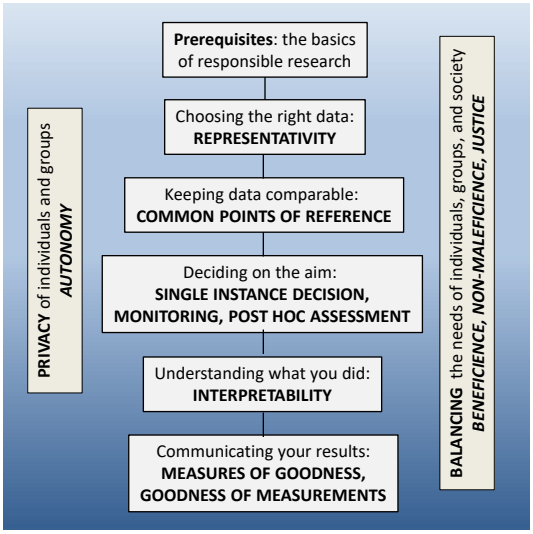


Fig. 6. Good Practice in Computational Paralinguistics

### 4.1 Prerequisites: The Basics of Responsible Research

A short remark on the fundamentals of doing good science, and the basics of releasing software: General principles for doing good, ethically responsible research are taught in introductory courses for students and young researchers. As they should be taken for granted, we only mention them shortly: full account of data recruitment, no plagiarism (of other work or of own work), no authorship without substantive own contribution to the work. These general principles are described in [33], [114]. Guidelines for making available databases for the scientific community can be found, e. g., in [115]. At the other end of the processing chain, the apps generated, e. g., for monitoring speech pathologies, need be safe against hacking. This is a very important but rather technical aspect when the software has been released 'into the wild'. The same holds for all kinds of personal data stored in the Web [44], [109].

10. Floridi et al., e. g., attribute privacy to non-maleficence [113]. We put more weight on the autonomy of individuals and their right to decide even if their choice might be unfavourable for themselves.

### 4.2 Choosing the Right Data: Representativity

Per definition, small samples are never representative; even larger samples are not representative for 'human kind' but mostly for *Western, Educated, Industrialised, Rich, and Democratic (WEIRD)* societies [116] and for groups that are more likely to be used as reservoir for recruiting subjects [117]. Other cultures are broadly underrepresented. Moreover, large samples can be outdated when used for research – this is partly unavoidable because it takes a long time to collect them, cf. [118]: "Many NLP tools for English and German are based on manually annotated articles from the Wall Street Journal and Frankfurter Rundschau. The average readers of these two newspapers are middle-aged (55 and 47 years old, respectively), and the annotated articles are more than 20 years old by now." Language itself is mostly not prone to fast change and does not easily adapt to societal development. In the lexicon, adaptation is fastest (neologisms); examples are the stigmatisation of terms like mongolism or negro. Morphology (such as generic vs gender-specific pronouns [119]), however, changes very slowly. Unsupervised, automatic learning represents those stereotypes that are implicitly entailed in language and speech (skewed distributions) and thus materialise as prejudices: The world turns out to be mostly male and white [120].

There are two types of representativity for a sample: *population* and *phenomenon*. Big data wants to approach population by more or less random sampling and large samples. Representativity is a basic requirement; yet, alone it is not sufficient. It has to be counterbalanced in case it reinforces the undesirable status-quo (e. g., repeat offenders are more often black in present day statistics, therefore, black people are more often not allowed a day's leave.) Social class – and by that, the probability that a loan is repaid, might be correlated with linguistic markers. The interests of the bank have to be counterbalanced by the interests of society not to stigmatise unduly specific social groups. The same holds for gender matters. Often, the algorithm is then called 'racist' or 'sexist'; yet, it is not the algorithm but the world mirrored by the algorithm that is biased, and it is the job of the scientists and ultimately, of society, to take countermeasures.

Our results might be representative for a specific (our own) culture but not for human kind in general: A nice example is speaker overlap as an indicator of conflict [121], [122], [123]; Grezes et al. [124] employ speaker overlap as a single feature exceeding a baseline for conflict obtained with 6,373 features [125] by 3% absolute. Yet, this 'Anglo' style does not extend to other cultures: In the 'Latin' conversational style, overlap indicates interest rather than conflict; in some Asian cultures ('Oriental' style), overlap is impolite and generally avoided [122], [123]. All these are aspects of an implicit *linguistic imperialism* – hard to avoid fully but we should be aware of the problem.

Most of the time, the phenomenon we are interested in cannot be seen and modelled directly; we use substitutes (stand-in data, *proxies*) instead. The transfer from proxy to phenomenon is, e. g., more direct when we model non-native speech: Speech is non-native when it sounds non-native; there is only one step from non-native speech to

detect the native language. The transfer is less direct, e. g., in the case of regional accent as proxy for neighbourhood which itself is taken as proxy for differentiating between good and bad debtors. The less direct the transfer is, the less reliable might be the modelling. Moreover, there is an implicit *cultural imperialism* in taxonomies when proxies are related to classes, e. g., when bad debtors are attributed to specific neighbourhoods. In the same vein, when we annotate according to specific theories, we can find a *theoretical imperialism*: The discrete, *big six* emotions [126], for example, representing for some time the prevailing scientific paradigm [127], somehow prevented other (types of) emotions to be seen in the scientific discourse that are equally representative [128].

Underestimated is the risk of *confounding classes* that have not been modelled: Often, an atypical class – for instance, people with Parkinson's Disease – has to be told apart from a typical class – in this case, people with comparable characteristics without Parkinson's Disease but with the same language background, age distribution, and suchlike. When we want to monitor progress, we stick to the very same class that is already known; there are no confounding classes. However, when we aim at screening a population for risk of Parkinson's Disease, there are many confounding classes: people with diseases that can show similar speech characteristics – depression, Alzheimer, or speech pathologies, or people that are 'goats' in the terminology of [129] which cannot be modelled, for unknown reasons. We might call this the *Closed World Fallacy*: A classification performance obtained in experiments with a few controlled classes cannot be transferred to real life where inevitably, performance will be (much) lower.

*Truth-in-advertising* should be obeyed: Summarising statements such as 'we can detect/handle/classify X' are misleading; we have to explicitly point out how representative our database is, what this means re prejudices, gender equality, racism, and stigmatisation in general, how direct the transfer from proxy to phenomenon is, and whether and how we take countermeasures to missing representativity. In this regard, there is no exact measure but we can detail in a section on *caveats*. Corbett et al. [130] propose suitable risk estimates, instead of "(1) anti-classification, meaning that protected attributes – like race, gender, and their proxies – are not explicitly used to make decisions; (2) classification parity, meaning that common measures of predictive performance (e.g., false positive and false negative rates) are equal across groups defined by the protected attributes; and (3) calibration, meaning that conditional on risk estimates, outcomes are independent of protected attributes." Other fairness-enhancing interventions are discussed in [131]; a popular scientific account of discrimination and related topics in AI is given in [132].

### 4.3 Keeping Approaches Comparable: Common Points of Reference

A new approach should be assessed with as many *Points of Reference* (POR) as possible. By that we mean standard 'entities' such as databases, partitions, feature vectors, procedures, and performance measures as, for instance, used in the Computational Paralinguistics ChallengE (ComParE), organised at Interspeech since 2009 [133]. Only by providing strict comparability with all other things being equal, we somehow can assess whether a new approach really is competitive or even superior. A POR example could be a publicly available database with a partitioning that it easy to reproduce, a competitive but at the same time, well-understood ML procedure such as Support Vector Machines (SVM), and a straightforward performance measure such as *Unweighted Average Recall (UAR)*[11] or correlation coefficients. Only then, we can evaluate the gain a new approach offers.

When a new approach yields (very) good results, employing new procedures/features and new data, we simply cannot say whether the data are favourable or something else. Only when the approach is tested with standard databases, we will find out. And vice versa, new databases should be tested with standard procedures such as SVM and, e. g., a well established feature set such as openSMILE [137], to give some baseline to compare with. Moreover, we should evaluate whether our model works equally well not only for unseen speakers, but also for unseen material.[12]

The concept of POR extends *replicability* [140], [141] in experimental-empirical studies: We not only need strict replications – which are anyway conceived as being less attractive in the scientific community – but some sort of 'weak' replication along the lines of PORs. Strict replications account for reliability, weak replications account for variability and by that, ensure higher robustness. Whereas representativity focuses on data, replicability focuses on the comparability of methods as well.

It might be difficult to establish a voluntary 'culture of POR', i. e., to introduce strategies and databases for such comparisons. We have to make it as comfortable as possible, and less cumbersome as far as time to invest is concerned, e. g., by setting up open challenges.

### 4.4 Deciding on the Aim: Single Instance Decision, Monitoring, and post hoc Assessment

The assessment of performance in CP and in ASP in general is mostly 'context-free', i. e., the whereabouts of concrete applications are not considered or only in a very cursory way. We can be 'fair' to all classes we want to model by using UAR. Yet we can also aim at, e. g., a high rate of true positives at the cost of increasing the rate of false positives when we want to reduce the search space for later processing [134], or a low rate of false positives when we want to be sure that we can safely interpret our (true) positives. Moreover, we can try to personalise our models, i. e., to have a closer look at the performance of individuals. This will get more important if it comes to applications in real-life, see Section 5.2.

11. UAR instead of the usual *Weighted Average Recall* has been introduced as 'average of class-wise recognition rates' [134], to facilitate a comparison for skewed class distributions; it has been used as a standard measure in the Computational Paralinguistics Challenges at Interspeech since 2009 [133], [135]. It is fair towards sparse – i. e., seen from an ethical point of view, 'minority' – classes; moreover, chance level is known when the number of classes is known (50% for 2 classes, 33.3% for 3 classes, and so on). Note that UAR is sometimes called 'macro-average', see [136].

12. This is done implicitly when, e. g., doing cross-corpus classification [138] but has to be done explicitly when, e. g., assessing non-native speech [139].

The requirements on performance measures differ considerably: In court, the identification of a defendant has to be perfect; this invalidates lie detectors employing speech, especially as single and only means. In contrast, when we evaluate atypical (non-native or pathological) speech, we only have to approach the performance of an expert panel or of a single expert – if the benchmark is given by the situation that only one expert evaluates in daily practice. This amounts to 100 % correct in court vs a correlation of 0.7 to 0.8 or so for speech assessments [77].

At least, we have to tell apart the following two constellations with clear, prototypical instances or with smooth transitions between them: *Single instance decisions (with instantaneous reaction)* must not harm the individual we decide upon. As mentioned above, this rules out the use of lie detectors in court. Note that it does not suffice for instantaneous reaction that automatic error rate equals human error rate: We simply do not know whether automatic errors are of the same weight as human errors. This is due to the standard procedure: Normally, some – mostly human – annotation is trained; by that, we sort of hard-code wrong and correct labels, and the error rate only gives the frequency of errors but not whether the actual ones are more or less serious. To cope with this problem, we needed, e. g., labels with different weight, depending on their impact in a real life scenario. The other constellation is *Monitoring/Screening (with delayed reaction and/or global post hoc assessment)*. In a call centre scenario, we can immediately react to the recognised anger of a caller or we can monitor his/her emotions and delay such a reaction. The first use is a yes/no decision, the latter one poses less strict demands because we can 'price in' errors, e. g., by using weights or taking into account confidences, and by considering implications of erroneous decisions; as for a taxonomy of such applications, cf. [78], [142].

### 4.5 Understanding what you did: Interpretability

Amongst the nowadays prevailing ML procedures, the characteristics of especially DNNs is *data greediness* and *opaqueness* (i. e., missing transparency) [143]. Data greediness is a practical problem in CP because we cannot simply collect a huge number of items as is the case for ASR or pictures: Mostly, we have to annotate our data or use more or less well-suited proxies; we are often interested in atypical data and this means in turn that these data are sparse. Data greediness is a theoretical problem as well because it hampers generalisation and interpretation – in a way, the classes DNNs recognise are extensionally defined, not intensionally. Opaqueness hampers interpretation. A rather indirect way of demonstrating the opaqueness of DNNs is via adversarial examples (cf. section 3.4) – small manipulations of samples, not recognisable for humans, lead to a drastic drop in performance. Adversarial examples for ASR are discussed in [144], [145], [146], for CP applications, in [100]. A critical appraisal of DNNs can be found in [147]; interpretability and explainability of classifiers in general and DNNs in particular are addressed in [148], [149], [150], [151], [152], [153].

A specific type of opaqueness is the confusion of correlation with causation in AI: On the one hand, this is an elementary mistake pointed out in introductory courses in psychology and other fields; on the other hand, this is not uncommon in AI, when big data produce – even by chance – high correlations that do not indicate some causation. Often, some proxy is employed as 'real' indication, e. g., a specific style of appearance in user groups (social marker) is interpreted as real indicator of sexual orientation [64].

Missing interpretability/explainability of DNNs is not only a theoretical problem – it might turn out to be, together with privacy and other ethical issues, the most important problem AI will face in the future: In its report on 'Civil Law Rules on Robotics' from 27.1.2017, the European Parliament recommends a principle of transparency: "... it should always be possible to supply the rationale behind any decision taken with the aid of AI that can have a substantive impact on one or more persons' lives; considers that it must always be possible to reduce the AI system's computations to a form comprehensible by humans; ..." [154]. Goodman and Flaxman [155] detail the *right to explanation* claimed in the EU's General Data Protection Regulation (GDPR) for algorithmic decisions that impact a user: "Indeed, machine learning can reify existing patterns of discrimination – if they are found in the training dataset, then by design an accurate classifier will reproduce them. In this way, biased decisions are presented as the outcome of an 'objective' algorithm."

The problem of how to structure the *input* into the ML procedure might be solved by appropriate measures to be taken, such as: balancing the training set, employing appropriate measures such as UAR, controlling the (type of) meta-data that are used, or explicitly biasing the results, e. g., by using weights. Thus, we can structure – i. e., manipulate for the better or the worse – the *input* into the ML procedure and later on, when looking at the output, re-structure if needed; this aims at representativity and intrinsic biases and is possible even if the ML procedure is rather a black box as in the case of DNNs. The *output* of ML procedures – classification or regression/correlation results, i. e., overall quality, hits, misses, false alarms – should, however, be interpretable as well.[13] Only then, we can decide whether to employ an ML procedure – especially if it is fully automatic – in critical situations, e. g., decisions on humans with high impact, and especially when we not only aim at global but at single instance decisions. As long as the procedure itself is a black box, we can relate input to output and gauge with trial and error; yet, we cannot interpret or even explain the outcome.

Thus, interpretability can be part of data selection and inclusion of meta-data but can be aimed as well at explaining the features that trigger decisions in classification. Whereas (meta-)data selection can be seen as a sort of preprocessing, interpretation of features is at the very heart of paralinguistics [157]. Studies on CP often fall short of explaining and interpreting results – not only in the case of DNNs where it, so far, seems to be impossible to do that; but also in the case of more 'classic' ML procedures where different feature selections could be employed. This might be due as well to the spurious results of automatic feature selection procedures with surviving but rather opaque features.

13. Doshi-Velez and Kim [156] define interpretability as "... the ability to explain or to present in understandable terms to a human".

A good POR for interpretability might be an SVM (instead of or complementary to DNNs) together with suited feature selection procedures such as wrappers; cf. [8, 235ff]. In addition, *knowledge-based features* should be employed and assessed as for their performance [157], [158], [159]; this circumvents the problem that automatically selected features are often less interpretable. Note that as long as the sparse data problem in CP has not been solved, the performance of SVM might not be (much) worse than the one of DNN.[14]

To make DNNs interpretable is not yet an established procedure but a topic for research. Yet, we can employ the same procedures as described in section 4.2.

### 4.6 Communicating your Results: Measures of Goodness and Goodness of Measures

Rosenthal and Blanck [4] stress that science of high quality is likely to be more ethically defensible; one pivotal aspect is the goodness of measures that researchers are using as standard in their scientific sub-culture. There is a noteworthy difference between articles in phonetics, socio-/psycholinguistics, psychology/sociology, and clinical studies on the one hand and articles within ASP on the other hand: In the first sub-cultures, you find papers where only p-values – the decisive measure in *Null Hypothesis Testing (NHT)* – are reported, in the other sub-culture, papers where only classifications and/or correlations/regressions are reported. Yet, in both cases, if p-values are small enough (usual thresholds below 0.05 or 0.001) or classification/regression high enough, this merits publication and establishing some theory or model, or claiming a performance that could be used in some application.

Gigerenzer and Marewski [160] describe the history of statistical inference and demonstrate how NHT emerged in the mid 1950ies as institutionalised combination of conflicting theories (Fisher and Neyman-Pearson). NHT as 'ritual' prevails until today, although it has been criticised from its beginning [161], [162]; cf. [160], [163], [164] and references therein. To mention some of the problems connected with NHT: p-values are generally not well understood [165]; they do not tell you that your hypothesis is true but that it is unlikely to get such a result, given the null hypothesis; they depend on sample size: large samples can yield significant results even if the differences are tiny; assumptions such as normal distribution and random sampling are seldom met; strict thresholds between 'true' and 'false' are nonsensical and in practice, rather employed as entrance ticket for publication. Twenty years ago, the American Psychological Association (APA) recommended to report effect sizes [166], and eventually, the American Statistical Association (ASA) published a statement on p-values [167] summarising all these problems – and of course, different assessments, concluding: "Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning." Further suggestions are given in [168].

Thus, instead of (only) reporting p-values, parameter estimates with confidence intervals and especially effect size measures should be reported, following sound exploratory statistics [169]. Effect size measures such as Cohen's $d$ can be mapped onto each other [170], [171], and onto the probability " ... that you could guess which group a person was in from knowledge of their 'score'." [170]. This last measure connects to the standard measure of classification (in this case, to a 2-class problem and UAR). Note that measures such as UAR give an overall picture; they do not tell us whether there are *sheep*, i. e., subjects that can be modelled to a high extent, or *goats*, i. e., subjects that cannot be modelled at all [129], in our sample; to find out, we have to have a look at the distribution across speakers.

Statistical measures should be understood across scientific sub-cultures; moreover, they should be understood by laypersons, i. e., in the societal discourse, as well. To this aim, McGraw and Wong proposed *Common Language Measures* that express "... how often a score sampled from one distribution will be greater than a score sampled from another distribution ..." [172]. In the same vein, Gigerenzer et al. [164] propose to describe results not in conditional probabilities but in natural frequencies – a measure that, e. g., is often used in societal discussions on the risks of breast or prostate cancer, or the spread of the coronavirus pandemic.

Scientific paradigms are like ocean liners: persistent and slow to change directions. Thus we cannot expect that phonetics and clinical studies abolish NHT instantaneously – cf. the recommendation of the APA to employ effect sizes in 1999 [166] and [171] where it is reported that more than ten years later, only half of the papers reported effect sizes. A step-by-step plan could look like this: First, as already called for by APA in 1999, effect sizes should be given alongside p-values. Second, argumentation should be based on effect size measures and not on p-values. CP studies might employ NHT – sometimes, this is asked for by reviewers and editors anyway. Effect sizes should be given that can be understood easily; even if Cohen's $d$ is well established, it is not fully optimal because its range of values is not confined to 0-1. Thus, best would to transform performance measures to countable entities [164] which are equally easily understood by other scientific sub-cultures and by society at large, by that laying a sound basis for the public discourse. Note that 'scientifically approved' measures need not be abandoned but complemented by suited common language measures.

## 5 Balancing the Needs

### 5.1 Basic Research

We cannot expect that the rules of good practice described in Section 4 can always be followed in a strict sense: Full *representativity* is almost impossible to obtain; yet, we have to be aware, should not make too far reaching conclusions, and

14. Decision trees are even more interpretable but in our experience, they fall short in terms of performance. Random Forests are better but have lost the transparency of single decision trees they exist of. Even if sometimes, feature interpretation is conducted, error analysis is seldom found; this might be due to the higher effort needed and/or to missing expertise.

should point out our limitations. There are no strict *points of reference* when we introduce a new database. However, we always can reason about our *aim* and about the appropriateness of our measures for different aims. So far, we can relate these three rules to *justice* in the sense of Principalism. The last two rules – *interpretability* and *goodness of measures* – rather relate to *autonomy*: The 'layman' in his/her role as user or patient should not only give *informed consent* but be able to make *informed decisions*. Overarching are of course the other rules of applied ethics: *beneficience/non-maleficience* and the *golden rule*, cf. Fig. 1.

Besides taking care of ethically responsible research, preserving privacy is the most important challenge. An easy way out of the dilemma not to give away personal information – including speech data – could simply be not to give away any data, be this *object data* (speech samples) or *meta data*. This would mirror the 'old days', some decades ago, where such a data exchange was not usual, due to practical problems (data were stored on tape and transferred to or mapped onto graphical representations such as spectrograms) and due to the wish to keep one's own 'property'. This strategy has changed for the better, and both raw data and extracted parameters are more and more claimed to be 'public property' – especially in the case of projects that have been publicly funded. This is based on the idea that resources should not be wasted. It should be added that both in science and marketing, claims are made that cannot be validated when the raw data are not available.

The classic approach not to give away information was continued in big companies such as Amazon/Apple/Google/Facebook/Microsoft and Samsung/Huawei who only partly release data or core algorithms. Yet, the research departments of these big companies at least take part in the scientific discourse, publish their work, and make some tools available. New and small enterprises, especially within CP, normally do not release data or algorithms, and by that, their claims cannot be scrutinised but only criticised, based on expert knowledge. Claims as the following are taken at face value: Voice prints are as good as finger prints, the lie detector works, we can diagnose your personality, we can predict whether you will develop Parkinson's Disease – and all this only by analysing your speech.

The only possible countermeasure is to make available data, meta data, and algorithms, e. g., in open challenges such as the Interspeech Computational Paralinguistics Challenges ComParE[15], MediaEval[16], CLEF[17], or AVEC[18]. However, this automatically means to go over to less strict privacy politics because of course, there are more possibilities to violate privacy if more people have access to the data. This dilemma cannot be solved but only balanced. In a free society, we cannot ban research, and the public wants us to do research for the good of society and individuals. Moreover, at least as far as speech is concerned, we have not seen (yet!) doxing, based *only* on speech. Of course, this is different if additionally, video information is available; cf. above Section 3.4.

Other fields, e. g., clinical studies, have much more elaborated demands on informations provided in publications such as conflicts of interest or documentation of data, and an established tradition of meta studies.[19] Such traditions could be taken as blueprints for research on CP as well.

### 5.2 Applications

In basic research, we normally model groups (sub-samples) out of a larger population, and present some performance measure; we do not have to decide. In applications – exemplified in the use cases described in Section 3.2, however, we have to make decisions on individuals or groups based on the more or less adequate performance obtained. *Common points of reference* are not in the fore, and *overall representativity* is no longer targeted but has to be granted, as a prerequisite for a successful – and ethically responsible – performance. *Interpretation* is getting more important: In court, an expert has to motivate and explain his/her results, e. g., when they claim to have identified a defendant based on speech recordings. The same will hold for health screening: Public and parliament have to get explanations when, e. g., new screening methods should be established and financed. A good *common language measure* that can be communicated and understood by the public is substantial. Most important might be the complex relationship between *single instance decisions* – mostly on individuals – and those relevant for groups (*monitoring/screening*) and eventually the one party that uses such applications: It is always an individual that gets accepted or rejected as candidate after personality assessment or after deciding upon regional accents as indicator of neighbourhood. Thus, we have to aim at a (personal) *individual representativity*.

In the long run, good or bad – and this means at the same time, ethically good or bad – decisions will impact the company or the agency that employ the tools as well.

## 6 Concluding Remarks

For practitioners, ethics is a thorny topic. This might explain why there is a large body of work on general ethics within philosophy but not that many studies on specific fields such as CP, targeting concrete steps to be taken. The concepts we introduced for good practice in Section 4 are not unique to CP; they have been addressed in the general (ethical) discourse as well – especially representativity and interpretability – and they are or should be topics within *(big) data science* [173] and neighbouring fields such as affective computing. We described them as a sequence of rules to be followed and exemplified with specific aspects of and use cases within CP where they have not yet received the attention they deserve.

Reasoning on ethics is nowadays often confined to threats to privacy and how to avoid that, especially with ECs. In this contribution, we wanted as well to shed some light onto other aspects that are equally important. They mostly relate to good practice – and by that, quality of science – and should be addressed in introductory courses as well. Yet, experience tells us that this is either not done

15. http://www.compare.openaudio.eu/
16. http://www.multimediaeval.org/
17. http://www.clef-initiative.eu/
18. https://sites.google.com/view/avec2018/home
19. http://www.cochrane.org/.

properly or it is still worth while to be pointed out. Research strategies are fossilised somehow, as well as scientific paradigms; yet, there should be some common goals, no matter how we want to achieve them, such as a common language measure that is understood by the public, and in the long run, interpretability of results. As far as privacy is concerned, matters are different: Privacy is being discussed at all societal levels and especially addressed in ECs. For the individual, a full preservation of privacy seems to be optimal; yet, there are societal reasons – and by that, reasons that will be manifest for individuals as well – for balancing the rules of privacy against the needs of society. This does not mean to undermine unduly the rules of privacy; it means, however, that we have to gauge benefits and risks for both individuals and society.

> *"... the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others."* John Stuart Mill, *On Liberty*, 1859.

## ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreements No. 338164 (ERC Starting Grant iHEARu), No. 688835 (RIA DE-ENIGMA), and No. 826506 (sustAGE). We are grateful to the editor and the anonymous reviewers for valuable suggestions.

## REFERENCES

[1] R. Cowie, "Ethical Issues in Affective Computing," in *The Oxford Handbook of Affective Computing*, R. Calvo, S. D'Mello, J. Gratch, and A. Kappa, Eds. Oxford: Oxford University Press, 2015, doi: http://dx.doi.org/10.1093/oxfordhb/9780199942237.013.006.

[2] C. Reynolds and R. Picard, "Affective sensors, privacy, and ethical contracts," in *CHI'04 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2004, pp. 1103–1106.

[3] A. Eriksson and F. Lacerda, "Charlatanry in forensic speech science: A problem to be taken seriously," *Journal of Speech, Language and the Law*, vol. 14, pp. 169–193, 2007.

[4] R. Rosenthal and P. D. Blanck, "Science and Ethics in Conducting, Analyzing, and Reporting Social Science Research: Implications for Social Scientists, Judges, and Lawyers," *Indiana Law Journal*, vol. 68, pp. 1209–1228, 1993.

[5] D. B. Resnik, *The Ethics of Science: An Introduction*. London and New York: Routledge, 1998.

[6] M. Iaccarino, "Science and ethics," *EMBO reports*, vol. 2, pp. 747–760, 2001.

[7] A. Batliner and B. Schuller, "More Than Fifty Years of Speech Processing – The Rise of Computational Paralinguistics and Ethical Demands," in *Proc. of ETHICOMP 2014*, Paris, France, June 2014, no pagination.

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

[9] "ethics," in *Encyclopedia Britannica: Micropedia*, Chicago, 1992, vol. 4, pp. 578–579, 15th edition.

[10] J. Fieser, "Ethics," https://www.iep.utm.edu/ethics/, 2019, retrieved 08/05/2020.

[11] L. Alexander and M. Moore, "Deontological ethics," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2016.

[12] W. Sinnott-Armstrong, "Consequentialism," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2019.

[13] R. Hursthouse and G. Pettigrove, "Virtue ethics," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2018.

[14] M. Sander-Staudt, "Care Ethics," in *The Internet Encyclopedia of Philosophy*, J. Fieser and B. Dowden, Eds., 2020, no pagination.

[15] J. Dittmer, "Applied ethics," https://www.iep.utm.edu/ap-ethic/, 2019, retrieved 13/05/2020.

[16] E. R. Winkler, "Applied Ethics, Overview," in *Encyclopedia of Applied Ethics*, R. Chadwick, Ed. London: Academic Press, Elsevier, 1998, pp. 191–196.

[17] F. Allhoff, "What Are Applied Ethics?" *Sci Eng Ethics*, vol. 17, pp. 1–19, 2011.

[18] T. L. Beauchamp and J. F. Childress, *Principles of biomedical ethics*. New York, NY: Oxford University Press, 2001.

[19] S. Döring, P. Goldie, and S. McGuinness, "Principalism: A Method for the Ethics of Emotion-Oriented Machines," in *Emotion-Oriented Systems: The Humaine Handbook*, R. Cowie, C. Pelachaud, and P. Petta, Eds. Berlin, Heidelberg: Springer, 2011, pp. 713–724.

[20] T. L. Beauchamp and D. DeGrazia, "Principles and Principlism," in *Handbook of Bioethics*, G. Khushf, Ed. Alphen aan den Rijn, NL: Kluwer Academic Publishers, 2004, pp. 55–74.

[21] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, pp. 389–399, 2019.

[22] J. Christman, "Autonomy in moral and political philosophy," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2018.

[23] T. Bynum, "Computer and information ethics," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2018.

[24] E. A. Buchanan and M. Zimmer, "Internet research ethics," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2018.

[25] M. Anderson and S. L. Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent," *AI Magazine*, vol. 18, pp. 15–26, 2007.

[26] I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson, N. R. Jennings, E. Kamar, I. M. Kloumann, H. Larochelle, D. Lazer, R. McElreath, A. Mislove, D. C. Parkes, A. S. Pentland, M. E. Roberts, A. Shariff, J. B. Tenenbaum, and M. Wellman, "Machine behaviour," *Nature*, vol. 568, pp. 477–486, 2019.

[27] E. Vayena and J. Tasioulas, "The dynamics of big data and human rights: the case of scientific research," *Phil.Trans. R. Soc. A*, vol. 374: 20160129, 2016, 14 pages.

[28] R. Herschel and V. M. Miori, "Ethics & Big Data," *Technology in Society*, vol. 49, pp. 31–36, 2017.

[29] L. Floridi and M. Taddeo, "What is data ethics?" *Phil.Trans. R. Soc. A*, vol. 374: 20160360, 2016, 5 pages.

[30] A. Zwitter, "Big data ethics," *Big Data & Society*, vol. 1, 2014, doi: 10.1177/2053951714559253.

[31] P. A. E. Brey, "Anticipatory Ethics for Emerging Technologies," *Nanoethics*, vol. 6, pp. 1–13, 2012.

[32] J. van den Hoven, M. Blaauw, W. Pieters, and M. Warnier, "Privacy and information technology," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2016.

[33] "European Textbook on Ethics in Research," 2010, retrieved 13/05/2020 from: https://ec.europa.eu/research/science-society/document_library/pdf_06/textbook-on-ethics-report_en.pdf.

[34] J. R. Herkert, "Future directions in engineering ethics research: microethics, macroethics and the role of professional societies," *Science and Engineering Ethics*, vol. 7, pp. 403–414, 2001.

[35] E. W. Schienke, S. D. Baum, N. Tuana, K. J. Davis, and K. Keller, "Intrinsic ethics regarding integrated assessment models for climate management," *Science and engineering ethics*, vol. 17, pp. 503–523, 2011.

[36] N. Elias, *Über den Prozeß der Zivilisation. Soziogenetische und psychogenetische Untersuchungen. Band 1: Wandlungen des Verhaltens in den weltlichen Oberschichten des Abendlandes / Band 2: Wandlungen der Gesellschaft: Entwurf zu einer Theorie der Zivilisation*. Basel: Verlag Haus zum Falken, 1939, english Translation: The Civilizing Process. Vol. 1: The History of Manners. Vol 2: State Formation and Civilization. Revised Edition, 2000. Basil Blackwell, Oxford.

[37] D. Webb, *Privacy and Solitude in the Middle Ages*. London: Hambledon Continuum, 2007.

[38] J. Holvast, "History of Privacy," in *The Future of Identity in the Information Society. Privacy and Identity 2008. IFIP Advances in Information and Communication Technology, vol 298*, V. Matyáš, S. Fischer-Hübner, D. Cvršek, and Švenda P., Eds. Springer, Berlin, Heidelberg, 2009.

[39] S. D. Warren and L. Brandeis, "The Right to Privacy," *Harvard Law Review*, vol. 4, pp. 193–220, 1890.
[40] "The American Psychological Association (APA) code of ethics, including 2010 and 2016 amendments," 2016, retrieved 13/05/2020 from http://www.apa.org/ethics/code/.
[41] "Association for Computing Machinery (ACM) Code of Ethics and Professional Conduct," 1992, retrieved 13/05/2020 from: https://www.acm.org/about-acm/acm-code-of-ethics-and-professional-conduct.
[42] "Charter of Fundamental Rights of the European Union," *Official Journal of the European Union*, vol. C83, pp. 389 – 403, 2010.
[43] "REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," Official Journal of the European Union, pp. 1–88, 2016, retrieved 13/05/2020 from: http://eur-lex.europa.eu/eli/reg/2016/679/oj.
[44] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps Towards a Common Understanding," in *Proc. of INTERSPEECH*, Graz, 2019, pp. 3695–3699.
[45] T. Hagendorff, "The Ethics of AI Ethics: An Evaluation of Guidelines," *Minds & Machines*, vol. 30, pp. 99–120, 2020.
[46] A. Stacey and J. Stacey, " Integrating Sustainable Development into Research Ethics Protocols," *The Electronic Journal of Business Research Methods*, vol. 10, pp. 54–63, 2012.
[47] T. Bond, "Ethical guidelines for researching counselling and psychotherapy," *Counselling and Psychotherapy Research*, vol. 4, pp. 10–19, 2004.
[48] V. Morrow and M. Richards, "The ethics of social research with children: An overview," *Children & society*, vol. 10, pp. 90–105, 1996.
[49] American Speech-Language-Hearing Association, "Code of Ethics," 2015, retrieved 13/05/2020 from: http://www.asha.org/Code-of-Ethics/.
[50] R. Body and L. McAllister, Eds., *Ethics in Speech and Language Therapy*. Chichester, UK: Wiley, 2009.
[51] D. L. Irwin, M. Pannbacker, T. W. Powell, and G. T. Vekovius, *Ethics for Speech-Language Pathologists and Audiologists: An Illustrative Casebook*. Clifton Park, NY, USA: Delmar Cengage Learning, 2007.
[52] Z. M. Schrag, "The Case against ethics review in the social sciences," *Research Ethics*, vol. 7, pp. 120–131, 2011.
[53] S. Plattner, "Comment on IRB regulation of ethnographic research," *American Ethnologist*, vol. 33, pp. 525–528, 2006.
[54] W. C. van den Hoonaard and A. Hamilton, Eds., *The Ethics Rupture: Exploring Alternatives to Formal Research-Ethics Review*. Toronto, Ontario, Canada: University of Toronto Press, Scholarly Publishing Division, 2016.
[55] W. C. van den Hoonaard and M. Tolich, "The New Brunswick Declaration of Research Ethics: A Simple and Radical Perspective," *Canadian Journal of Sociology*, vol. 39, pp. 87–97, 2014.
[56] R. Cowie, "Editorial: 'Ethics and Good Practice' – Computers and Forbidden Places: Where Machines May and May Not Go," in *Emotion-Oriented Systems: The Humaine Handbook*, R. Cowie, C. Pelachaud, and P. Petta, Eds. Berlin, Heidelberg: Springer, 2011, pp. 707–711.
[57] ——, "The good our field can hope to do, the harm it should avoid," *IEEE Transactions on Affective Computing*, vol. 3, pp. 410–423, 2012.
[58] P. Goldie, S. Döring, and R. Cowie, "The Ethical Distinctiveness of Emotion-Oriented Technology: Four Long-Term Issues," in *Emotion-Oriented Systems: The Humaine Handbook*, R. Cowie, C. Pelachaud, and P. Petta, Eds. Berlin, Heidelberg: Springer, 2011, pp. 725–733.
[59] I. Sneddon, P. Goldie, and P. Petta, "Ethics in Emotion-Oriented Systems: The Challenges for an Ethics Committee," in *Emotion-Oriented Systems: The Humaine Handbook*, R. Cowie, C. Pelachaud, and P. Petta, Eds. Berlin, Heidelberg: Springer, 2011, pp. 753–767.
[60] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, H. Anderson, H. Roff, G. C. Allen, J. Steinhardt, C. Flynn, S. Ó. hÉigeartaigh, S. Beard, H. Belfield, S. Farquhar, C. Lyle, R. Crootof, O. Evans, M. Page, J. Bryson, R. Yampolskiy, and D. Amodei, "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," arXiv:1802.07228, 2018.
[61] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.
[62] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *PNAS*, vol. 110, pp. 5802–5805, 2013.
[63] Y. Wang and M. Kosinski, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images," *Journal of Personality and Social Psychology*, vol. 114, pp. 246–257, 2018.
[64] A. Gelman, G. Mattson, and D. Simpson, "Gaydar and the fallacy of objective measurement," *Sociological Science*, vol. 5, pp. 270–280, 2018.
[65] O. Bendel, "The Uncanny Return of Physiognomy," in *AI and Society: Ethics, Safety and Trustworthiness in Intelligent Agents, AAAI 2018 Spring Symposium Series*, PaloAlto, CA, 2018, pp. 10–17.
[66] R. P. Gaudio, "Sounding Gay: Pitch Properties in the Speech of Gay and Straight Men," *American Speech*, vol. 69, pp. 30–57, 1994.
[67] S. E. Linville, "Acoustic Correlates of Perceived versus Actual Sexual Orientation in Men's Speech," *Folia Phoniatr Logop*, vol. 50, pp. 35–48, 1998.
[68] D. Rendall, P. L. Vasey, and J. McKenzie, "The Queen's English: An Alternative, Biosocial Hypothesis for the Distinctive Features of "Gay Speech"," *Arch Sex Behav*, vol. 37, pp. 188–204, 2008.
[69] L. Zimman, "Hegemonic masculinity and the variability of gay-sounding speech – The perceived sexuality of transgender men," *Journal of Language and Sexuality*, vol. 2, pp. 1–39, 2013.
[70] A. Hanani, M. Russell, and M. Carey, "Human and computer recognition of regional accents and ethnic groups from British English speech," *Computer Speech and Language*, vol. 27, pp. 59–74, 2013.
[71] S. Fazel, Z. Chang, T. Fanshawe, N. Långström, P. Lichtenstein, H. Larsson, and S. Mallett, "Prediction of violent reoffending on release from prison: derivation and external validation of a scalable tool," *The Lancet Psychiatry*, vol. 3, pp. 535 – 543, 2016.
[72] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 Speaker Trait Challenge," in *Proc. of INTERSPEECH*, Portland, OR, 2012, pp. 254–257.
[73] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, pp. 273–291, 2014.
[74] M. Raghavan, S. Barocas, J. M. Kleinberg, and K. Levy, "Mitigating Bias in Algorithmic Employment Screening: Evaluating Claims and Practices," *CoRR*, vol. abs/1906.09208, 2019. [Online]. Available: http://arxiv.org/abs/1906.09208
[75] B. Schuller, E. Marchi, S. Baron-Cohen, H. O'Reilly, D. Pigat, P. Robinson, I. Davies, O. Golan, S. Fridenson, S. Tal, S. Newman, N. Meir, R. Shillo, A. Camurri, S. Piana, A. Staglianò, S. Bölte, D. Lundqvist, S. Berggren, A. Baranger, and N. Sullings, "The state of play of ASC-Inclusion: An Integrated Internet-Based Environment for Social Inclusion of Children with Autism Spectrum Conditions," in *Proc. of 2nd International Workshop on Digital Games for Empowerment and Inclusion (IDGEI 2014)*, Haifa, Israel, 2014, 8 pages.
[76] C. C. Ragin and L. M. Amoroso, *Constructing social research: The unity and diversity of method*. Sage publications, 2010.
[77] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "Peaks – a system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, pp. 425–437, 2009.
[78] A. Batliner, M. Neumann, F. Burkhardt, A. Baird, N. T. Vu, and B. Schuller, "Ethical Awareness in Speech Emotion Processing: Taxonomies for Applications," ms., submitted, 2020.
[79] S. Furui, "Fifty years of progress in speech and speaker recognition," *The Journal of the Acoustical Society of America*, vol. 116, pp. 2497–2498, 2004.
[80] ——, "40 years of progress in automatic speaker recognition," in *Proc. of Advances in Biometrics: Third International Conference, ICB 2009, Alghero, Italy, June 2-5*, M. Tistarelli and M. S. Nixon, Eds. Berlin, Heidelberg: Springer, 2009, pp. 1050–1059.
[81] K. Sparck Jones, "Natural language processing: an overview," in *International encyclopedia of linguistics*, W. Bright, Ed. New York: Oxford University Press, 1992, pp. 53–59.

[82] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," *Journal of the American Medical Informatics Association*, vol. 18, pp. 544–551, 2011.
[83] T. W. Smith, "Changing Racial Labels: From "Colored" to "Negro" to "Black" to "African American"," *The Public Opinion Quarterly*, vol. 56, pp. 496–514, 1992.
[84] D. Hovy and S. L. Spruit, "The Social Impact of Natural Language Processing," in *Proc. of 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, 2016, pp. 591–598.
[85] S. Barocas and H. Nissenbaum, "Big data's end run around procedural privacy protections," *Communications of the ACM*, vol. 57, pp. 31–33, 2014.
[86] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," in *IEEE Symposium on Security and Privacy*, Oakland, CA, 2008, 15 pages.
[87] ——, "Privacy and security Myths and fallacies of 'Personally identifiable information'," *Communications of the ACM*, vol. 53, pp. 24–26, 2010.
[88] "REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016," 2016, retrieved 13/05/2020 from: http://data.europa.eu/eli/reg/2016/679/oj.
[89] "Handbook on European data protection law," 2014, retrieved 13/05/2020 from: https://fra.europa.eu/en/publication/2018/handbook-european-data-protection-law.
[90] D. M. Douglas, "Doxing: a conceptual analysis," *Ethics Inf Technol*, vol. 18, pp. 199–210, 2016.
[91] M. R. Jacobs, "Institutional Review Boards and Independent Ethics Committees," in *Principles of Good Clinical Practice*, M. J. McGraw, A. N. George, S. P. Shearn, R. L. Hall, and J. Thomas F. Haws, Eds. London: Pharmatceutical Press, 2010, pp. 121–147.
[92] M. Lindorff, "Ethics, ethical human research and human research ethics committees," *Australian Universities' Review*, vol. 52, pp. 51–59, 2010.
[93] W. H. Organisation, "Research ethics committees. Basic concepts for capacity-building," Printed by the WHO Document Production Services, Geneva, CH, 2009.
[94] R. D. L. C. Bernabe, G. J. M. W. van Thiel, J. A. M. Raaijmakers, and J. J. M. van Delden, "The risk-benefit task of research ethics committees: An evaluation of current approaches and the need to incorporate decision studies methods," *BMC Medical Ethics*, vol. 13, 2012, 9 pages.
[95] W. D and M. FG, "Assessing research risks systematically: the net risks test," *Journal of Medical Ethics*, vol. 33, pp. 481–486, 2007.
[96] C. Glackin, G. Chollet, N. Dugan, N. Cannings, J. Wall, S. Tahir, I. G. Ray, and M. Rajarajan, "Privacy Preserving Encrypted Phonetic Search of Speech Data," in *Proc. of ICASSP*, New Orleans, 2017, pp. 6414–6418.
[97] P. Lopez-Otero, C. Magarios, L. Docio-Fernandez, E. Rodriguez-Banga, D. Erro, and C. Garcia-Mateo, "Influence of speaker de-identification in depression detection," *IET Signal Processing*, vol. 11, pp. 1023–1030, 2017.
[98] J. M. S. Dias, A. Abad, and I. Trancoso, "Exploring Hashing and Cryptonet Based Approaches for Privacy-Preserving Speech Emotion Recognition," in *Proc. of ICASSP*, Calgary, AB, 2018, pp. 2057–2061.
[99] F. C. Teixeira, A. Abad, and I. Trancoso, "Privacy-preserving Paralinguistic Tasks," in *Proc. of ICASSP*, Brighton, UK, 2019, pp. 6575–6579.
[100] Y. Gong and C. Poellabauer, "Crafting Adversarial Examples For Computational Paralinguistic Applications," arXiv:1711.03280, 2017.
[101] N. Carlini and D. Wagner, "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text," arXiv:1801.01944v1, 2018.
[102] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion?" in *Proc. of INTERSPEECH*, Graz, 2019, pp. 3700–3704.
[103] D. Leroy, A. Coucke, T. Lavril, T. Gisselbrecht, and J. Dureau, "Federated learning for keyword spotting," in *Proc. of ICASSP*, 2019, pp. 6341–6345.
[104] J. Konecný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated Optimization: Distributed Machine Learning for On-Device Intelligence," *CoRR*, vol. abs/1610.02527, 2016. [Online]. Available: http://arxiv.org/abs/1610.02527
[105] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How To Backdoor Federated Learning," *CoRR*, vol. abs/1807.00459, 2018. [Online]. Available: http://arxiv.org/abs/1807.00459
[106] Yücel Saygin and Dilek Z. Hakkani-Tür and Gökhan Tür, "Sanitization and Anonymization of Document Repositories," in *Database Technologies: Concepts, Methodologies, Tools, and Applications*, J. Erickson, Ed. Information Science Reference, 2009, pp. 2129–2139.
[107] G. Beigi, K. Shu, R. Guo, S. Wang, and H. Liu, "I am not what i write: Privacy preserving text representation learning," *ArXiv*, vol. abs/1907.03189, 2019.
[108] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.
[109] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, M. A. Abdelraheem, A. Abad, F. Teixeira, D. Matrouf, M. Gomez-Barrero, D. Petrovska-Delacrétaz, G. Chollet, N. Evans, T. Schneider, J.-F. Bonastre, B. Raj, I. Trancoso, and C. Busch, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441–480, 2019.
[110] J. L. Kröger, O. H.-M. Lutz, and P. Raschke, "Privacy Implications of Voice and Speech Analysis – Information Disclosure by Inference," in *IFIP International Summer School on Privacy and Identity Management*. Springer, 2019, pp. 242–258.
[111] F. Pokorny, K. Bartl-Pokorny, C. Einspieler, D. Zhang, R. Vollmann, S. Bölte, M. Gugatschka, B. Schuller, and P. Marschik, "Typical vs. atypical: Combining auditory Gestalt perception and acoustic analysis of early vocalisations in Rett syndrome," *Res Dev Disabil.*, vol. 82, pp. 109–119, 2018.
[112] F. B. Pokorny, P. B. Marschik, C. Einspieler, and B. Schuller, "Does She Speak RTT? Towards an Earlier Identification of Rett Syndrome Through Intelligent Pre-Linguistic Vocalisation Analysis," in *Proc. of INTERSPEECH*, San Francisco, CA, 2016, pp. 1953–1957.
[113] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and Machines*, vol. 28, pp. 689–707, 2018.
[114] Carolyn Bertozzi et al., *On Being a Scientist: A Guide to Responsible Conduct in Research: Third Edition.*, National Academy of Sciences, National Academy of Engineering (US) and Institute of Medicine (US) Committee on Science, Engineering, and Public Policy, Ed. Washington (DC): National Academies Press (US), 2009.
[115] F. Schönbrodt, M. Gollwitzer, and A. Abele-Brehm, "Der Umgang mit Forschungsdaten im Fach Psychologie: Konkretisierung der DFG-Leitlinien," *Psychologische Rundschau*, vol. 68, pp. 20–35, 2017.
[116] J. Henrich, S. J. Heine, and A. Norenzayan, "The weirdest people in the world?" *The Behavioral and brain sciences*, vol. 33, pp. 61–83; discussion 83–135, 2010.
[117] R. G. Smart, "Subject selection bias in psychological research," *Canadian Psychologist*, vol. 7a, pp. 115–121, 1966.
[118] D. Hovy and A. Søgaard, "Tagging Performance Correlates with Author Age," in *Proc. of the 53rd Annual Meeting of ACL-IJCNLP*, Beijing, China, 2015, pp. 483–488.
[119] A. Batliner, "The comprehension of grammatical and natural gender: a cross-linguistic experiment," *Linguistics*, vol. 22, pp. 831–856, 1984.
[120] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, pp. 183–186, 2017.
[121] K. Hilton, "The Perception of Overlapping Speech: Effects of Speaker Prosody and Listener Attitudes," in *Proc. of INTERSPEECH*, San Francisco, CA, 2016, pp. 1260–1264.
[122] F. Trompenaars and C. Hampden-Turner, *Riding the Waves of Culture: Understanding Diversity in Global Business*, 2nd ed. McGraw-Hill Companies, Incorporated, 1998.
[123] H. Fitzgerald, *How Different are We? Spoken Discourse in Intercultural Communication*. Clevendon, UK: Multilingual Matters, 2003.
[124] F. Grèzes, J. Richards, and A. Rosenberg, "Let me finish: automatic conflict detection using speaker overlap," in *Proc. of INTERSPEECH*, Lyon, 2013, pp. 200–204.

[125] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. of INTERSPEECH*, Lyon, 2013, pp. 148–152.

[126] P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*, T. Dalgleish and M. Power, Eds. New York: John Wiley, 1999, pp. 301–320.

[127] T. S. Kuhn, *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 1962.

[128] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements," *Psychological Science in the Public Interest*, vol. 20, pp. 1–68, 2019.

[129] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "SHEEP, GOATS, LAMBS and WOLVES. A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation," in *Proc. of ICSLP*, Sydney, 1998, no pagination.

[130] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *CoRR*, vol. abs/1808.00023, 2018. [Online]. Available: http://arxiv.org/abs/1808.00023

[131] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 2019, pp. 329–338.

[132] K. Zweig, *Ein Algorithmus hat kein Taktgefühl: Wo künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können*. München: Heyne, 2019.

[133] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. of INTERSPEECH*, Brighton, 2009, pp. 312–315.

[134] A. Batliner, R. Kompe, A. Kießling, M. Mast, H. Niemann, and E. Nöth, "M = Syntax + Prosody: A syntactic–prosodic labelling scheme for large spontaneous speech databases," *Speech Communication*, vol. 25, pp. 193–222, September 1998.

[135] A. Rosenberg, "Classifying skewed data: Importance weighting to optimize average recall," in *Proc. of INTERSPEECH*, Portland, OR, 2012, pp. 2242–2245.

[136] C. D. Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2009.

[137] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. of the 21st ACM International Conference on Multimedia*, ser. MM '13, New York, NY, USA, 2013, pp. 835–838. [Online]. Available: http://doi.acm.org/10.1145/2502081.2502224

[138] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, "Cross-Corpus Classification of Realistic Emotions – Some Pilot Experiments," in *Proc. of 3rd International Workshop on EMOTION: Corpora for Research on Emotion and Affect, satellite of LREC 2010*, Valletta, Malta, 2010, pp. 77–82.

[139] F. Hönig, A. Batliner, and E. Nöth, "Automatic assessment of non-native prosody – annotation, modelling and evaluation," in *Proc. of IS ADEPT, International Symposium on Automatic Detection of Errors in Pronunciation Training, Stockholm, Sweden, June*, 2012, pp. 21–30.

[140] J. P. A. Ioannidis, "Why Most Published Research Findings Are False," *PLOS Medicine*, vol. 2, pp. 696–701, 2005. [Online]. Available: https://doi.org/10.1371/journal.pmed.0020124

[141] R. Moonesinghe, M. J. Khoury, and A. C. J. W. Janssens, "Most published research findings are false – but a little replication goes a long way," *PLOS Medicine*, vol. 4, pp. 218–221, 2007.

[142] A. Batliner, F. Burkhardt, M. van Ballegooy, and E. Nöth, "A Taxonomy of Applications that Utilize Emotional Awareness," in *Proceedings of IS-LTC 2006*, Ljubliana, 2006, pp. 246–250.

[143] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Attentive Explanations: Justifying Decisions and Pointing to the Evidence (Extended Abstract)," arXiv:1711.07373v1, 2017.

[144] M. Cisse, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling Deep Structured Prediction Models," arXiv:1707.05373v1, 2017.

[145] L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding," arXiv:1808.05665v1, 2018.

[146] M. Alzantot, B. Balaji, and M. Srivastava, "Did you hear that? Adversarial Examples Against Automatic Speech Recognition," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017, 6 pages.

[147] G. Marcus, "Deep learning: A critical appraisal," *CoRR*, vol. abs/1801.00631, 2018. [Online]. Available: http://arxiv.org/abs/1801.00631

[148] Z. C. Lipton, "The mythos of model interpretability," *CoRR*, vol. abs/1606.03490, 2016. [Online]. Available: http://arxiv.org/abs/1606.03490

[149] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," *CoRR*, vol. abs/1602.04938, 2016. [Online]. Available: http://arxiv.org/abs/1602.04938

[150] W. Samek, T. Wiegand, and K.-R. Mller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," arXiv:arXiv:1708.08296v1, 2017.

[151] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[152] B. D. Mittelstadt, C. Russell, and S. Wachter, "Explaining Explanations in AI," *CoRR*, vol. abs/1811.01439, 2018. [Online]. Available: http://arxiv.org/abs/1811.01439

[153] C. Molnar, "Interpretable machine learning," 2019, https://christophm.github.io/interpretable-ml-book/.

[154] "REPORT with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))," 2017, retrieved 13/05/2020 from: https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html.

[155] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a "right to explanation"," arXiv:1606.08813v3, 2016.

[156] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," arXiv:1702.08608v2, 2017.

[157] A. Batliner and B. Möbius, "Prosody in Automatic Speech Processing," in *The Oxford Handbook of Language Prosody*, C. Gussenhoven and A. Chen, Eds. Oxford, UK: Oxford University Press, 2020, 20 pages, to appear.

[158] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, and J. Krajewski, "Acoustic-Prosodic Characteristics of Sleepy Speech – between Performance and Interpretation," in *Proc. of Speech Prosody 2014*, Dublin, 2014, pp. 864–868.

[159] M. P. Black, D. Bone, Z. I. Skordilis, R. Gupta, W. Xia, P. Papadopoulos, S. N. Chakravarthula, B. Xiao, M. V. Segbroeck, J. Kim, P. G. Georgiou, and S. S. Narayanan, "Automated evaluation of non-native English pronunciation quality: combining knowledge- and data-driven features at multiple time scales," in *Proc. of INTERSPEECH*, Dresden, Germany, 2015, pp. 493–497.

[160] G. Gigerenzer and J. N. Marewski, "Surrogate Science: The Idol of a Universal Method for Scientific Inference," *Journal of Management*, vol. 41, pp. 421–440, 2015.

[161] W. Rozeboom, "The Fallacy of the Null-Hypothesis Significance Test," *Psychological Bulletin*, vol. 57, pp. 416–428, 1960.

[162] H. Eysenck, "The Concept of Statistical Significance and the Controversy about one-Tailed Tests," *Psychological Review*, vol. 67, pp. 269–271, 1960.

[163] G. Gigerenzer, "Mindless statistics," *The Journal of Socio-Economics*, vol. 33, pp. 587–606, 2004.

[164] G. Gigerenzer, W. Gaissmaier, E. Kurz-Milcke, L. M. Schwartz, and S. Woloshin, " Helping Doctors and Patients Make Sense of Health Statistics," *Association for Psychological Science*, vol. 8, pp. 53–96, 2008.

[165] L. Badenes-Ribera, D. Frías-Navarro, H. M. i Bort, and M. Pascual-Soler, "Interpretation of the p-value: A National Survey Study in Academic Psychologists from Spain," *Psicotema*, vol. 27, pp. 290–295, 2015.

[166] L. Wilkinson, "Statistical methods in psychology journals: Guidelines and explanations," *American Psychologist*, vol. 54, pp. 594–604, 1999.

[167] R. L. Wasserstein and N. A. Lazar, "The ASA's statement on p-values: context, process, and purpose," *The American Statistician*, vol. 70, pp. 129–133, 2016.

[168] R. L. Wasserstein, A. L. Schirm, and N. A. Lazar, "Moving to a world beyond "p $<$ 0.05"," *The American Statistician*, vol. 73, pp. 1–19, 2019.
[169] J. W. Tukey, *Exploratory Data Analysis*. Reading, Mass. Menlo Park, Cal., London, Amsterdam, Don Mills, Ontario, Sydney: Addison-Wesley Publishing Company, 1977.
[170] R. Coe, "It's the effect size, stupid: What effect size is and why it is important," Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, 12-14 September, 2002, retrieved 02/16/2019 from: www.leeds.ac.uk/educol/documents/00002182.htm.
[171] C. O. Fritz, P. E. Morris, , and J. J. Richler, "Effect size estimates: Current use, calculations, and interpretation," *Journal of Experimental Psychology: General*, vol. 141, pp. 2–18, 2012.
[172] K. O. McGraw and S. Wong, "A common language effect size statistic," *Psychological Bulletin*, vol. 111, no. 2, p. 361, 1992.
[173] V. Dhar, "Data Science and Prediction," *Communications of the ASC*, vol. 56, pp. 64–73, 2013.

**Anton Batliner** received his doctoral degree in Phonetics in 1978 at LMU Munich. He is now with the Chair of Embedded Intelligence for Health Care and Wellbeing at University of Augsburg, Germany. He is co-editor/author of two books and author/co-author of more than 300 technical articles, with an h-index of $>$45 and $>$10000 citations. His main research interests are all (cross-linguistic) aspects of prosody and (computational) paralinguistics.

**Simone Hantke** received her PhD in 2018 from the Technische Universität München (TUM), Germany. She is now with audEERING GmbH. The topic of her doctoral thesis is data collection and new machine learning approaches for robust automatic speech recognition and speaker characterisation. She is the lead author of the crowdsourcing platform iHEARu-PLAY.

**Björn Schuller** received his doctoral degree in 2006 and his habilitation in 2012, all in electrical engineering and information technology from TUM in Munich, Germany. He is Professor of Artificial Intelligence in the Department of Computing at the Imperial College London/UK, Full Professor and head of the Chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg/Germany, and CSO of audEERING – an Audio Intelligence company. He (co-)authored 6 books and more than 900 publications in peer reviewed books, journals, and conference proceedings leading to more than 30 000 citations (h-index = 80).