Ethical awareness in paralinguistics: a taxonomy of applications

Anton Batliner, Michael Neumann, Felix Burkhardt, Alice Baird, Sarina Meyer, Ngoc Thang Vu, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Batliner, Anton, Michael Neumann, Felix Burkhardt, Alice Baird, Sarina Meyer, Ngoc Thang Vu, and Björn W. Schuller. 2023. “Ethical awareness in paralinguistics: a taxonomy of applications.” *International Journal of Human–Computer Interaction* 39 (9): 1904–21. https://doi.org/10.1080/10447318.2022.2140385.

Ethical Awareness in Paralinguistics: A Taxonomy of Applications

Anton Batliner[a], Michael Neumann[b,f], Felix Burkhardt[c,e], Alice Baird[a], Sarina Meyer[b], Ngoc Thang Vu[b] and Björn Schuller[a,d]

[a] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany; [b] Institute for Natural Language Processing (IMS), University of Stuttgart, Germany; [c]audEERING GmbH, Germany; [d]GLAM Group on Language, Audio, and Music, Imperial College London, UK; [e]Technical University of Berlin, Germany; [f]Modality.AI, Inc., San Francisco, US

ARTICLE HISTORY
Compiled November 2, 2022

ABSTRACT

Since the end of the last century, the automatic processing of paralinguistics has been investigated widely and put into practice in many applications, both on wearables, the smartphone, and on the computer (home and host). In this contribution, we address ethical awareness for paralinguistic applications, by establishing taxonomies for data representations, system designs for and a typology of applications, and users/test sets and subject areas. These are related to an 'ethical grid' consisting of the most relevant ethical cornerstones, based on principalism. The characteristics of and the interdependencies between these taxonomies are described and exemplified. This makes it possible to assess more or less critical 'ethical constellations'. To the best of our knowledge, this is the first attempt of its kind.

Index Terms: ethics, paralinguistics, taxonomies, speech emotion processing, applications

1. Introduction

Whereof one cannot speak, thereof one must be silent. Ludwig Wittgenstein

The *subject* of this contribution is *ethical awareness* of applications by design and by reasoning; the *field* we take our use cases from is (computational) *paralinguistics* (CP), see Schuller and Batliner (2014); the *tools* we employ for describing and assessing applications are independent but highly related *taxonomies* for data representation, ethical cornerstones, system design, a typology of applications, and users and subject areas. These tools make available terms and inter-dependencies – on a higher level, they enable us to speak about the important aspects of ethical awareness; in practice,

This is an original manuscript (preprint) of an article submitted in June 2021 to the International Journal of Human-Computer Interaction, Taylor & Francis, and eventually accepted for publication in October 2022. The published article will be available at https://doi.org/10.1080/10447318.2022.2140385. Changes in the accepted manuscript are minor and relate mainly to grammar and style. Figures and tables can be found at the appropriate places and not at the end of the manuscript.

CONTACT Michael Neumann. Email: michael.neumann@modality.ai

it is a list of aspects to be ticked as applicable or not, the same way as we tick off a list with things to take with us when travelling – not really necessary for the 'inaugurated' but highly welcome and mandatory for the layman and in case we might forget – ethics is a topic widely discussed, yet at the same time easily forgotten in daily practice. Combinations of these taxonomies constitute three different – but again highly related – loops for doing research, implementing applications, and taking care of ethical awareness.

In this section, we will now introduce these topics one by one and then motivate the added value of integrating them – which we want to elaborate on in the remaining sections. Sec. 2 presents and discusses the taxonomies: R: Representations of data; E: Ethical cornerstones; S: System design, T: Typology of applications; U: Users and subject area; and P: Principalism. Sec. 3 establishes a typology of applications within CP. Sec. 4 elaborates on the interaction and interdependencies between the taxonomies established. The remaining sections shortly address the problems ethical awareness faces – ethics washing in Sec. 5 and general caveats in Sec. 6.

1.1. Ethical Awareness

The great theories of ethics – virtue, deontological, and utilitarian ethics, together with other ethical theories, constitute the 'backbone' of – but have to be adapted to and put into practice within – the field of applied ethics with its *golden rule* (Cowie, 2012) and especially the principles of **beneficence**, **non-maleficence**, **autonomy**, and **justice**[2], characterised as **principalism** (Döring, Goldie, & McGuinness, 2011); an ethical framework taking into account these principles for AI – referring to them as principles within bioethics – can be found in Floridi et al. (2018). An overview on the ethics of computing is given in Stahl, Timmermans, and Mittelstadt (2016). Jobin, Ienca, and Vayena (2019) surveyed existing ethical guidelines and added, as amongst the most frequently used, the principles of **transparency, fairness, responsibility, privacy, freedom, trust, dignity, sustainability**, and **solidarity**; see as well Fjeld, Achten, Hilligoss, Nagy, and Srikumar (2020); Leslie (2019); Lo Piano (2020). We will refer to these terms as covering and specifying principalism; this is done in a loose, sort of 'family resemblance' fashion, without trying to establish an extended applied ethical theory. *Ethical awareness* is understood as taking care of those aspects of ethics that are relevant for one's practice.[3] Benjamins, Barbado, and Sierra (2019) proposed measures for *Responsible AI by Design*, which are partly similar to the ones we propose; whereas they focus on implementing ethical principles in large organisations, we focus on implementing and taking care of these principles when designing applications in one specific field, namely CP.

1.2. Paralinguistics

We define paralinguistics along the lines of Batliner, Hantke, and Schuller (2020); Schuller and Batliner (2014); it covers three vocal and verbal aspects: [+vocal/+verbal], modulated onto or entailed in speech, e. g., prosody; [+vocal,-verbal], embedded within speech, e. g., laughter, screams; and [-vocal,+verbal], (spoken or) written text – conno-

[2] Boldface denotes all those terms that are displayed in Fig. 1 and Fig. 2, and italics indicates other important terms.

[3] Note that we are not concerned with implementing ethical theories in machines (Tolmeijer, Kneer, Sarasua, Christen, & Bernstein, 2020) but with systematically assessing ethical awareness in applications.

tations of words and phrases; it does not cover [-vocal, -verbal], i. e., facial gestures, gait, posture, and every other context, which can be attributed to the field of *Affective Computing* (AC), see Picard (1997). There is some overlap between CP and AC; yet, to give two examples, facial gestures are not dealt with in CP, and non-native speech as speaker characteristic is part of CP but not of AC.

Arguably, most research has been done on emotion processing; early studies date back to the 1990s, e. g., on automatic speech emotion recognition (Dellaert, Polzin, & Waibel, 1996) and emotion synthesis (Schröder, 2001). Yet, this stands prototypically for other paralinguistic (and affective) phenomena such as mood, personality, and all types of typical and atypical (e. g., pathological) phenomena and variants; an overview and a sketch of the historic developments are given in Schuller and Batliner (2014). *Speech Emotion Processing* (SEP) encompasses generation (language), synthesis (acoustics), and recognition of emotion in speech and language. Since then, we have seen a plethora of approaches towards collecting data, modelling emotions and other paralinguistic phenomena, benchmarking with challenges, and developing applications that utilise *emotional awareness*. Emotion is a fuzzy term; in a prototypical use and in specific theories, it is confined to a few (four, six, or a bit more) *basic emotions*; in everyday use and in *affect* theories, less clear-cut states and traits are encompassed, e. g., interest, boredom, and frustration, and in *personality* theories, traits like the *big five* are modelled.

According to some records, the "emotion-detection and -recognition market was worth $12 billion in 2018, and by one enthusiastic estimate, the industry is projected to grow to over $90 billion by 2024" (Crawford et al., 2019). In the public discourse, attention now focuses on (missing) *ethical awareness*. This is mirrored in the scientific community by studies that challenge basic assumptions (Batliner et al., 2020) and performance claimed, cf. Barrett, Adolphs, Marsella, Martinez, and Pollak (2019) for "inferring emotions from human facial movements". We do not know of any similar large-scale study addressing the same topic in SEP or CP in general. This cannot be given in this contribution; instead, we want to introduce some ideas towards taxonomies for applications in the field; by that, we take up again and extend the taxonomy addressed in Batliner, Burkhardt, van Ballegooy, and Nöth (2006) and discuss the most important 'building bricks' of representation of states and traits and their ethical implications for real life applications.

1.3. The Term 'Taxonomy'

In the Encyclopedia Britannica ("Taxonomy", 1992), the definition reads as follows: "*Taxonomy*, in a broad sense, the science of classification, but more strictly the classification of [...] organisms [...]" In biology, taxonomy constitutes the foundation of this science, see C. (1926) and the problem that trivial conceptual errors in biological taxonomies can have severe consequences (Bortolus, 2008); it was pointed out that the neglect of taxonomic work might simply be due to its neglect by the present day publication system (Samyn & Massin, 2002). Not as prominent but in a similar way, taxonomies were in the past in the focus of (historical) linguistics but gave way for more universalistic, experimental, or theoretical approaches. We use the term 'taxonomy' because it is not yet employed by theories/approaches within our fields such as 'class' (classification), trait (personality), dimension (emotion modelling) – just to mention a few candidates. Note that we do not want to evaluate ethical behaviour such as honesty in taxonomies as in Ghahari et al. (2010) but to present intrinsic characteristics and interdependencies in taxonomies.[4]

1.4. Motivation

Johnson and Wetmore state: “[...] nearly every decision an engineer makes is not simply a detached technical decision but has ethical and value content and implications” (Johnson & Wetmore, 2008). It makes sense to pigeonhole such decisions – resulting in a taxonomy or several taxonomies – to be able not only to evaluate each single decision but to evaluate their interdependencies as well. Value by design (Friedman, 1996) and bias in computer systems (Friedman & Nissenbaum, 1996) have already been discussed in the 1990s. During the last years, the intrinsic bias in AI due to (not only) skewed distributions in training corpora (especially racism and sexism) has been much debated. As prerequisites of countermeasures, *data sheets* (Gebru et al., 2018) or *data statements* have been proposed: “A data statement is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software” (Bender & Friedman, 2018). This might seem utopic but would be the best way to account for “... exclusion, overgeneralization, and underexposure ... [and] ... generalizability and reproducibility.” (Bender & Friedman, 2018). The counterpart for machine learning (ML) models are *model cards*: “Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (e. g., race, geographic location, sex, ...)” (Mitchell et al., 2018); see as well Holstein, Vaughan, Daumé III, Dudík, and Wallach (2018) who propose means how to take into account the requirements of practitioners in ML systems. On the one hand, data sheets and model cards protocol ‘objective’ data such as distribution of classes in corpora or decision thresholds; on the other hand, all these are of course moving targets that have to be constantly calibrated, the same way as the ICD (International Classification of Diseases) (World Health Organization (WHO), 1993) modifies criteria for defining diseases.

1.5. A Note on Terminology

Terminology is notoriously fuzzy both within AI and within ethics: There are many competing theories of and practical approaches towards ethics, each of them not necessarily with identical terms; the field of AI did rather evolve in a partly parallel, partly sequential fashion. A nice example is the ‘evolution’ of the terms *interpretability* and *explainability* in AI; they are not easily told apart (Mittelstadt, Russell, & Wachter, 2019). Arrieta et al. (2020) display them as keywords (in titles, abstracts, keyword lists) used in articles during the last decade; the use of ‘Interpretable AI’ increased but seems to be decreasing now, while ‘Explainable AI’ is gaining more attention. Yet, there are no clear cut (intensional) definitions of these two terms; they are either used in a loose fashion, or attempts to tell them apart are not fully convincing. We will thus resort to an extensional definition which seems to mirror a common use: We reserve ‘interpretation’ to the processes inside the algorithmic box, and ‘explanation’ to the input into and the output out of the box. However, this can only be done for prototypical constellations.

[4] Our taxonomies are more concrete than those in Chancellor, Birnbaum, Caine, Silenzio, and Choudhury (2019) which as well could be called ‘aspects of’.

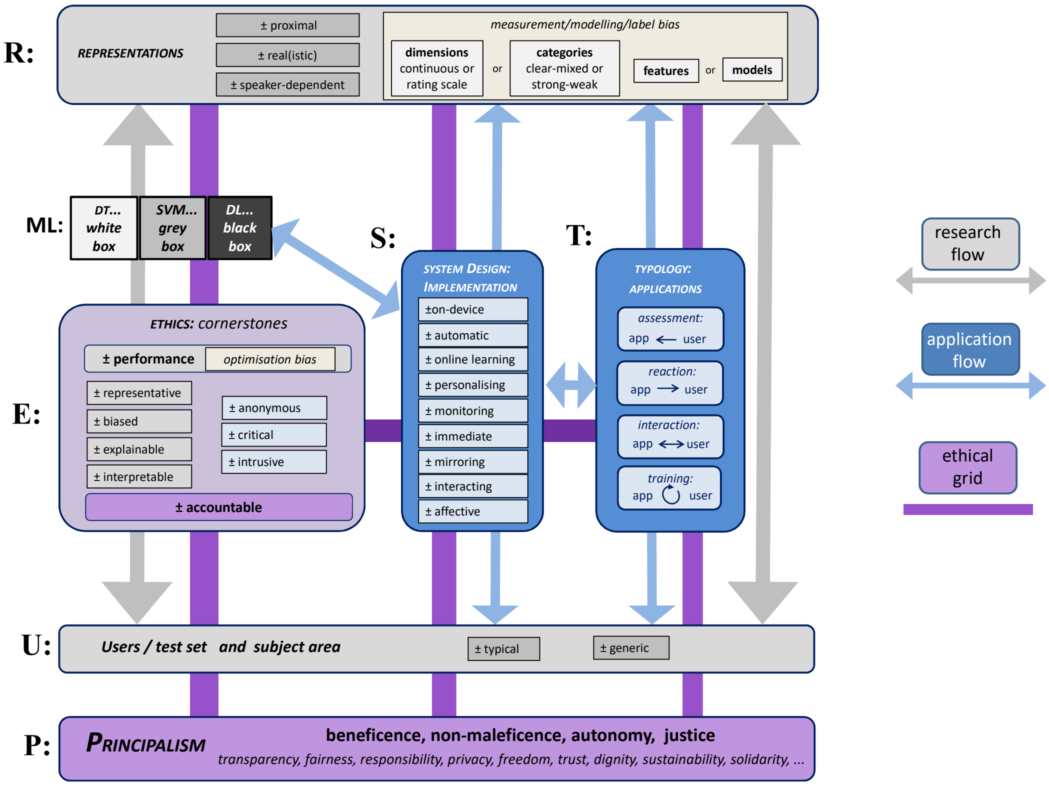

Figure 1. Taxonomies for paralinguistic applications: R: **R**epresentations of data; E: **E**thical cornerstones; S: **S**ystem design, T: **T**ypology of applications; U: **U**sers and subject area; P: **P**rincipalism; ML: different types of Machine Learning procedures.

As a consequence, when attempting to describe approaches and concepts, people resort to collecting uses of terms and display them as semantic fields (*word clouds*, as the one for explainable AI, or XAI, in Adadi and Berrada (2018)) instead of trying to establish strict definitions and hierarchies of terms. In turn, we try to keep our definitions and terms consistent within our own approach but cannot relate each term to other approaches or theories. Most importantly, we do not want to use those terms within our taxonomies that are frequently employed within ethics theories, as documented in Jobin et al. (2019); for instance, 'transparency' is employed within ethics, thus, we substitute it with a near-synonymous term, namely 'accountability', in **E** (ethical cornerstones), see Sec. 2.

2. Taxonomies

In Fig. 1 we summarise five taxonomies: **R:** data and how they are represented, discussed in Sec. 2.1; **E:** ethical cornerstones, explained in Sec. 2.2; **S:** system design, i. e., characteristics of the implementation with more or less impact on ethics, see Sec. 2.3; **T:** a typology of applications, given in Sec. 3; and **U:** the different types of users and subject areas/phenomena dealt with, see Sec. 2.4. They are all related to and evaluated with regard to **P:**, principalism, i. e., the underlying ethical principles, see Sec. 1.1. There are basically three process flows depicted in Fig. 1: (i) the *research flow*, indicated by taxonomies with grey background and by grey boxes and arrows. For that, we need representations of the real world (**R**) as input into our algorithmic processing – indicated by the small boxes termed ML – that has to meet the requirements described in the grey boxes of (**E**) and is evaluated on unseen test data – here, represented by **U** and iteratively improved with new data from **R**; (ii) the *application flow*, connecting the two blue boxes with each other and with input and output (blue arrows), describing **R:** data used for training, **S:** the implementation principles, **T:** the typology of applications, and **U:** the users who employ the application; and (iii) the *ethical grid*, indicated by violet bars connecting all taxonomies with the focus on ethical awareness, i. e., **E** and **P**; the grey boxes in **E** denote scientific aspects, and the blue ones application specific ones; **P**, dark violet, depicts the overarching ethical principles.

The first two process flows can be seen as 'void of ethics' – not in the sense of 'being unethical per se' but in the sense of simply not necessarily and not always taking into account ethical awareness – they would work without any regress to ethics; yet, our topic is exactly the connection between **P** and **E** (as specifications of **P**) and the other taxonomies. These are the relationships of and interdependencies between the taxonomies: **R** and **U**: real world and proxies thereof as input into the machinery and target of output; **U** guides the design of **T**; **S** fleshes out **T**; **E** assesses **R**, **S**, and **T**, and takes care of **U**; **P** is reference for **E**, where the other taxonomies are evaluated.

ML is, on the one hand, at the heart of the matter; yet, we only want to address it from the point of view of ethics – not with a full taxonomy but as 'processing box' – especially from those aspects depicted in **E**, i. e., as defined in Sec. 1.5, explanation for input/output (**R**/**U**), and interpretation for the box: Decision trees (DT) exemplarily stand for procedures that are rather fully transparent (white box), Support Vector Machines (SVM) for those that are in between (grey box), and Deep Learning (DL) procedures for those that are rather fully opaque (black box). For SVMs, we can for instance employ acoustic features and wrappers that are computationally costly because a model is tested for each (subset of) features, but they normally yield highly competitive performance (Batliner et al., 2011) and can be interpreted. Other methods are, e. g., based on correlation or information gain (Schuller & Batliner, 2014, p.235 ff.). Such procedures would be theoretically possible for DL as well but they are, in practice, hardly feasible because of time constraints. Strictly speaking, when using a wrapper, we do not really look inside the box as we do when evaluating a DT; yet, it is qualitatively much more than in the case of DL, where we at most evaluate proxies as input (**R**) and compare them with the output (**U**). An overview of techniques for XAI is given in Adadi and Berrada (2018).

We want to establish taxonomies for applications and their prerequisites that are mostly conceived as dimensions, having endpoints denoted as plus and minus; they have either no values in between – then, they are purely binary – or values at some intermediate stages as well meaning 'more or less'. Endpoints are, e. g., processing 100 % on device vs 100 % in the cloud ([± **on-device**]); more or less distributed processing is located somewhere between these endpoints. These dimensions are not intended to be definite, but to be a basis for discussion and further development: Researchers, developers, and customers can assess applications according to their characteristics as more or less 'ethically aware', based not only on common sense but on an exhaustive list of criteria and interdependencies.

2.1. Representations: Data and how they are Represented

Fig. 1 **R** gives an overview of the representational aspects.

[± **proximal**]: *Stand-ins (Proxies) can more or less represent the phenomenon we are interested in.*
The object data are represented in acoustic and/or written form, and mapped onto some symbolic representation of *proxies*. This is done by utilising metadata and/or by annotations. For example, for credit scoring, we can assess the credit history of the costumer, or their neighbourhood; the first is more proximal than the latter. When we try to recognise heart rate as ground truth with the help of speech data, beats per minute are fully proximal as sole target, and rather proximal as indicator for stress.

[+ **real(istic)**]: *Representations of paralinguistic content and features/models are only **representative** when the data they are obtained from are representative.*
Normally, we (should) aim at real life data, i. e., [+ **real(istic)**]. Too often, data are still *acted*, i. e., [− **real(istic)**], and/or drawn from an unrepresentative and by that, **biased** sample (one culture, one language, one ethnicity/gender/age group (Barrett et al., 2019; Elfenbein & Ambady, 2003) – to mention just a few factors).

[± **speaker-dependent**]: *Normally, **speaker-independence** is aimed at.*
This is good for addressing new data, but sub-optimal when we want to model known users (personalisation) – especially important for health care, and when we want to explain and interpret the outcome. Further aspects are detailed in Sec. 2.3.

The representation can either be *dimensional* or *categorical*. Arguably, most used in emotion modelling are the two dimensions *arousal* and *valence*, and the big n categories such as anger, sadness, joy, and some other 'basic' ones. This is the 'traditional' approach that, as far as we can see, has been implemented in most of the applications. Yet, in research papers, other dimensions and categories are discussed widely, especially whether emotions normally are *clear* and *strong* (such as the big n) or rather *mixed* and *weak*. The same way as for emotions, for all other paralinguistic phenomena, dimensions can be mapped onto a categorical (or ordinal) representation, and several categories can be located onto dimensions. We are still far from knowing which of these different (types of) representations are most adequate.

Whether 'classic' **features** (such as the ones from openSMILE (Eyben, Weninger, Groß, & Schuller, 2013)), that can be interpreted within phonetic/linguistic models, or less concrete (deep, end-to-end) **models** are employed, is foremost a matter of taste and eventually, a matter of performance or efficiency. So far, features are still competitive and widely used in applications; this might change when really big data are available for training. The basic *ethical question* is which representation can be harnessed for interpretation and explanation or interaction (tutoring) with users (Batliner & Möbius, 2020).

Dobbe, Dean, Gilbert, and Kohli (2018) introduced several types of *technical bias*: (1) **measurement bias**, when for example an ordinal scale is transformed into a binary nominal scale; (2) **modelling bias**: features are used as proxies – how representative are these? (3) **label bias**: do our labels denote 'historic' outcomes or likely proxies? (4) **optimisation bias**: do we, e. g., optimise UAR[5] or true positives? The first three types are part of data representations as depicted in Fig. 1 **R**. Optimisation bias will be further discussed in section 2.2. These (possible) biases have to be taken care of

[5] UAR stands for Unweighted Average Recall, i. e., the average of the values (True Positives in percent) in the diagonal of a confusion matrix, for $n = 2$ or more classes.

when we try to explain and interpret, see as well Sec. 2.2. Note that these technical biases cannot be avoided – but taken care of: Inevitably, measurement and labelling transform the data, and then models and optimisation procedures the outcome as well.

2.2. Ethics: Cornerstones

Fig. 1 **E** shows a taxonomy of pivotal criteria (cornerstones) for *ethical awareness*; these can be part of implementations or characteristics of presenting the application to the user. Grey boxes relate to Fig. 1 **R**, i. e., to decisions taken 'outside' of the application design. These ethical cornerstones cover the aspects consistently highlighted in the literature and which we propose to be relevant for speech-based applications in particular. We first present the cornerstones that are central for the scientific discourse, and then those that focus on the user (light blue background), and end up with accountability, covering all pivotal ethical aspects that have to be communicated to the user – and to society at large.

[± **performance**]: *High performance, meaning a system is* 'good enough' *for the use case vs low, insufficient performance.*

Performance is foremost and initially a number that denotes the quality/goodness of a system in terms of a chosen evaluation metric. For instance, let us assume that it is, for a binary classification task, a number between 50 (chance level) and 100 – expressed in percent and called "UAR"; it can be expressed in other ranges, or it can be verbalised (poor, good, ...). The assessment of performance can be based on: (i) *existence*: We have an algorithm that models X; (ii) *comparison*: we are better than chance (significantly) / earlier experiments / others have been / a challenge baseline; (iii) *extrapolation*: we will be better when we will have better AI and / or more data; see Batliner et al. (2020). In other fields, statistical significance is used as a threshold – if p-values are lower than a pre-defined threshold, the results are usually accepted; this is highly problematic and should be abandoned (Batliner et al., 2020; Wasserstein & Lazar, 2016). When aimed at and eventually put into practice, i. e., deployed in applications, it can be asked whether **performance** is *really* good enough for employing an application in 'real-life', i. e., [+**performance**]. It often turned out – albeit not often assessed – that algorithms have lower **performance** when applied on new, unseen data. Therefore, practitioners do prefer more generic approaches that are not only optimised for specific data, by that accepting lower **performance**. In the societal discourse, **performance** is weighted against the odds, thus, a vaccine with relatively low **performance** will be accepted when better alternatives are not available. We can mark cornerstones for the threshold of **performance** acceptance: 100 % correct is necessary for yes/no decisions in court; as low a **performance** as 70 % can be acceptable for assessing something when human experts are not better. Apart from that, **performance** is claimed in scientific publications and product marketing from companies but has to be conceived rather as a promise than a proven fact. The above-mentioned **optimisation bias** (Dobbe et al., 2018) is closely related to the discussion on **performance**. The important question to ask here is whether a model is optimised sensibly for the intended use case or for another metric in order to report high **performance**. High **performance** usually means that a system is optimised towards specific data and is therefore [+**biased**]; low(er) **performance** can of course mean bad modelling, but as well that features and/or ML procedures are less specific but more generic, i. e., [–**biased**]. We can optimise extremely for a specific sample; this means (i) we optimise as well for characteristics that rather have to do with speaker individual traits and not with the phenomenon we are interested in, or (ii) we optimise for the sub-population, e. g., Italian learners of English, or (iii) we try to optimise for the target language English. The possibility to interpret is different for (ii) and (iii), and is rather spurious for (i). **Performance** ties algorithms and models with the real world: These use proxies, see Fig. 1 **R**, but have to 'cope' with the real world behind these proxies, see Fig. 1 **U**. Although not as often addressed in debates on ethics as, e. g., privacy and bias, it plays a decisive role. We thus attribute **performance** to ethical cornerstones and not to implementation.

[± **representative**]: *Full representativity, in the sense of random selection out of a population – in the case of CP, mostly not extensionally defined – is almost never met; thus, we have to assume more or less representative samples.*

Different aspects in CP applications concern representation: subjects (speakers), data, and models. Most attention so far is given to subjects – to right or wrong sampling out of a population, as a consequence of selection-bias in ML, when discussing shortcomings of applications. However, how a phenomenon itself is modelled is equally important for representativity. In the case of emotion analysis, some researchers are calling for an overhaul on how emotion itself can be understood in computational analysis (Barrett et al., 2019). Current categorical approaches result in a mismatch between data and processing; by that, ethically questionable results are obtained. Of course, not only subjects, but (training) data have to be [+**representative**] as well. Representativity is not a matter of system design, but of decisions made 'outside', i. e., on (types of) more or less biased training data. [**-Representative**] systems are in contrast to the principle of **justice** and risk to be maleficent, as an abundance of controversial incidences in the ML community have highlighted, including model sexism, racism, and homophobia (Resnick, 2019). However, **fairness**-enhancing interventions are becoming more prevalent, in an effort to reduce discrimination and increase representation (Friedler et al., 2019).

[± **biased**]: *Model decisions are biased, as a consequence of the underlying data, vs this is not the case.*

At first sight, a **biased** algorithm is simply not representative. Yet, representativity is sort of nested: When the population is **biased** in the sense that classes are not equally distributed, then this bias is representative for this population; yet, it might be unfairly modelling the data against the minority classes. Thus, a bias can be introduced to counterbalance unwelcome consequences if data are fully representative (Danks & London, 2017). We can tell apart *distribution bias* given in representative (Phillips et al., 2009) data from *fairness bias*, i. e., skewed data introduced to guarantee fairness. Biases have a large impact on explainability and interpretability; therefore, they should be continuously monitored and not taken for granted. Bias problems can easily be tested by subdividing the test set into biased groups, as has been done for gender in Buolamwini and Gebru (2018). We have to tell apart classification tasks where we aim at a mapping of representativity from input to output from prescription tasks where we, e.g., assess job applications based on representative but biased training data. In the first type, ethics might not be in the fore, in the second, it certainly is.

[± **explainable**]: *Applications provide methods to analyse the model's behaviour based on input-output relations vs this is not possible.*

The classic approach towards explainability is to first observe the outcomes of the box – classifications, decisions, and alike. If these look biased, we then take a look at the input into the box; this means the different types of technical biases mentioned above, the choice of proxies, and especially the distribution of characteristics – gender, ethnicity, social class, severity of a condition, and alike – in the training data. We then can try to counter-balance biases seen in the output by calibrating the input distribution. When looking at a model's inputs and outputs to explain its decisions, we can tell apart: (1) *pseudo-proxies* such as modelling microphones or room acoustics instead of the targeted phenomena – these are errors and have to be avoided; (2) *partial proxies*, such as self or expert/other assessment[6]; this is unavoidable but they have to be used and interpreted carefully – they do not constitute a ground truth but at best a gold standard; (3) *full proxies*, e. g., a reliable biological signal as reference, i. e., a real ground truth (Stappen et al., 2021). A [+**explainable**] system provides mechanisms to uncover pseudo-proxies and to gain an understanding of what is actually modelled. One example for such methods are so-called heat-maps, which highlight those parts in the input signal (a picture or a speech signal, e. g., represented as a spectrogram) that are relevant for the model's decision and relate this to the phenomenon that is to be modeled. An attempt within speech processing can be found in Krug and Stober (2018).

[± **interpretable**]: *Applications provide methods to analyse and understand their internal information processing vs black box systems.*

Methods used in [+**interpretable**] applications should provide enough evidence to explain their predictions, e. g., perceive emotion A and B with different scores because of features X and Y, and disclose their internal decisions. As opposed to [±**explainable**], interpretability is concerned with 'opening and looking into the box'. Interpreting features is an established approach (Batliner & Möbius, 2020); within AI/ML, explaining techniques like LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro, Singh, & Guestrin, 2016) are proposed to overcome the black-box dilemma. This cornerstone is highly relevant to **transparency** and **trust**. Furthermore, [–**interpretable**] systems might jeopardise the principle of **justice** as they are usually not predictable.[7] As outlined in Y. Zhou and Danks (2020), different varieties of interpretability (or intelligibility) should be distinguished based on the targeted user group and their goals, e. g., engineers/researchers, users of an application, or others that are affected by the system. Interpretability – as we use the term – particularly addresses the research and developer community (as for many applications, understanding the inner workings is not relevant for users to achieve their goals). However, [+**interpretable**] systems can also provide useful information for affected persons or users when it comes to being accountable, as detailed below. In scientific papers, we can establish a coarse typology of approaches towards interpretability: (1) *no interpretation*, just performance; (2) *wrong interpretation* (combined with wrong 'explanation'): e. g., attribution to certain speech features when room acoustics are in fact the decisive factor; (2) *partial interpretation*: only relevant for specific cases; (3) *full interpretation*: fully relevant for *individual models* (personalisation) or *population models* (generic use). Further, we can tell apart causes and/or expert (knowledge-based) features most relevant for recognition/classification from those that are most relevant and can be harnessed for therapy and treatment, see the difference between *power features* that contribute to a high **performance** and *leverage features* that can be interpreted and conveyed, e. g., in training, to the user (Batliner & Möbius, 2020). Interpretation has much to do with error analysis, and this in turn with explainability. Note that **performance** and interpretability do not

[6] Self-assessment is usually taken as being closer to the ground truth; yet, all types of assessment are no real ground truth but filtered.

[7] EU Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act); https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence [Accessed May 07, 2021]

go together smoothly: When we optimise with automatic feature selection, we will surely find opaque features that cannot be interpreted; when we optimise less (e. g., by employing only expert features), we have some lower **performance** but can interpret better. When we train with a specific data set, then inevitably, we are not generic enough, but might get higher **performance**. Generic reach will go together with lower **performance**. All the attempts towards 'cross-modelling are so far not really successful because data are spurious. An approach towards interpretability is constraining the learnable input for neural nets by something that makes sense for the task at hand. For example, Ravanelli and Bengio (2019) introduced an architecture where the first layer of filters for a convolutional net are not learnt but designed to be bandpass filters for which only the low and high cutoff frequencies are learnt. Further discussions on interpretability can be found in (Doran, Schulz, & Besold, 2017; Doshi-Velez & Kim, 2017; Kim et al., 2018; Lipton, 2016; Molnar, Casalicchio, & Bischl, 2020; Y. Zhou & Danks, 2020).

We now describe the more user-centred ethical cornerstones that are less focal for the 'pure scientific discourse' – but highly relevant for ethics because they impact the user – and by that, society – to a high degree.

[± **anonymous**]: *No identifying or personal information is passed or processed vs personal data is processed and stored.*

A prominent regulation in this respect is the European General Data Protection Regulation (GDPR).[8] The main issues are that users keep control over their data – they can demand insight and deletion, and data is collected only to a minimal extent to fulfil the application's purpose. This becomes complex if the main goal of an application is to collect data that might, in the long run, be used to train systems that are beneficient. Often, data collectors try to prevent a maleficent use of the data by forbidding commercial use in the license; this leads to a situation where commercial systems cannot be compared with the ones developed by academics, and prevents beneficent use of data. As a compromise, **privacy** can be granted by pre-processing speech, e. g., pre-training models locally and passing only weights to update the algorithm, as with block-chain/federated learning approaches (Leroy, Coucke, Lavril, Gisselbrecht, & Dureau, 2019). Within the COMPRISE EU project (Tomashenko et al., 2020), a privacy preserving framework is introduced by anonymising training data on the user device. Interestingly, the EU project ECoWeB (Newbold et al., 2020) tackles the **privacy** problem with a contrasting strategy by explicitly not storing any data on the user device but keeping it on the server safely from malevolent attempts.

[± **critical**]: *The application's aims are impaired if the targeted phenomenon (e. g., emotion) is predicted erroneously vs erroneous processing does not impair an application's overall aims.*

[+**Critical**] applications cannot fulfill their use if the underlying model is not working correctly, for example if emotional states are misclassified or unconvincing emotional expressions are rendered. As this potentially harms a user who has been promised a working system and relies on it, [+**critical**] applications should pay special attention to the underlying model's performance. The fact that emotional AI is prone to a certain error rate should be made transparent to the user. In order to guarantee this, a certain confidence and explainability is indispensable. Of course, applications within the health domain are more critical than, e. g., applications for language learning. With respect to the principle of **justice**, being [+**representative**] for the intended

[8] `https://eur-lex.europa.eu/eli/reg/2016/679/oj` [Accessed April 22, 2021]

user group is of special importance for [+**critical**] applications.

[± **intrusive**]: *The application's aim can have serious consequences for the user's economic, physical, or mental well-being vs no such consequences.*

In contrast to [±**critical**], [±**intrusive**] is intentional. Serious consequences can be both negative and positive. In any case, [+**intrusive**] applications especially need to pay attention to ethical principles. Fully [–**intrusive**] are for instance pure fun, Tamagotchi-like applications (Cowie, 2012) – unless, of course, its use is getting obsessive.

[± **accountable**]: *A system's methods, internal decision processes and outcomes are transparent (and communicated to users) vs this is not the case.*

As a prerequisite, the user has to be (made) aware that their data are recorded, documented, and processed. The same way as HTTP cookies require some active confirmation from the users, the application has to account for the type of interaction and processing of personal information. In a first step, users have to be told *that* their data are processed; the next step is to disclose *how* their data are processed. Accountability is chosen as term for transferring the scientific discussion on explainability and interpretability into the societal discourse and especially into the design of applications. In general, A is **accountable** to B when A informs B about A's actions and decisions to justify them and to suffer punishment in the case of misconduct. In the AI context, the big questions are: 1) 'to which/whom can A be referred?' – A could be AI developers, designers or institutions; and 2) can AI systems be **accountable** in the way humans are, i. e., can AI systems be considered as A as a stand-alone system? Or should humans always be the only actors who are ultimately responsible for technological artefacts? Although clear answers to these questions are difficult to reach, we suggest [+**accountable**] applications to be transparent in: (a) methods how to extract paralinguistic information, e. g., which information sources are used, (b) mechanisms how the output is used in the system architecture, e. g., whether the dialogue policy is conditioned by paralinguistic information, and if yes, in which way, and (c) implementation how to use paralinguistic information in the decision making processes, e. g., in the context of recruitment. This could be the first tiny meaningful step to answer the first question. Furthermore to support the 'right to explanation' in the European GDPR, we suggest that [+**accountable**] applications should have the ability to communicate in an automatic way to the outside world about their internal processes by requests. Systems need to be able to answer the question 'why', explaining the role of paralinguistic information and beyond (e. g., personality) in the final decision making processes, e. g., 'why was a person rejected in a hiring process?'. Although humans, e. g., in the role of HR officers, are mostly avoiding this kind of explanation, systems should communicate their decisions. This cornerstone is relevant to **transparency**, **justice**, **fairness**, and **trust**.

2.3. System Design: Implementation

Fig. 1 **S** displays a taxonomy for types of implementation that basically – although not necessarily – are binary; we will exemplify these criteria in Sec. 3. We first present three basic decisions taken and then those types that characterise user-system interactions.

[±**on-device**]: *Processing is done completely on the user's device vs cloud-based processing.*

This design choice can impact **privacy** and is often also related to performance.

While [+**on-device**] processing guarantees that no data is shared with others, the performance of ML models that can run on user devices like smartphones is often very limited compared to large powerful models that are executed on a server. Cloud-based processing can still be [+**anonymous**], but all necessary precautions have to be taken.

[±**automatic**]**:** *System decisions and actions are fully automatic vs assisting human based decisions.*
See Torous and Roberts (2017) and human agency in RichardBenjamins (2021). Fully [+**automatic**] implies [+**critical**] because there is no human intervention possible. Depending on the degree of automation and the application context, explainability and accountability play an important role in [+**automatic**] systems.

[±**online learning**]**:** *A system learns autonomously and continuously as new data becomes available, also called continuous learning or self-learning, vs non-autonomously periodical model updates.*
See RichardBenjamins (2021). Especially for [+**personalising**] applications, continuous learning can be beneficial as the system can learn and adapt to the particular user with every new data point. From the ethical perspective, the question of accountability is crucial and at the same time even more difficult to answer for systems that learn autonomously. Further, such systems could unintentionally become [+**biased**] and [**-representative**] over time, depending on the incoming data; see the example of the chatbot Tay (Wolf, Miller, & Grodzinsky, 2017), discussed in section 3.

[±**personalising**]**:** *Application adapts to the user, based not only on meta-information but iteratively vs no personalisation.*
The system and the application based on it can be intended to be generic, i. e., (i) *population-based* – which is, however, not really possible even if confined to, e. g., speakers of one language; (ii) *sample-based*, i. e., aiming at a smaller and thus more specific group; or (iii) *individual-based*, i. e., personalised; the more generic, the less precise the system will be (Barnett & Torous, 2019; Dumont, Giergiczny, & Hess, 2015). Thus, personalised means less bias, population-based more bias. A sample can be open-set – then, the sample has to be representative for the population, or closed-set – then, the sample is exhaustively and extensionally defined.

[±**monitoring**]**:** *(Continuous) monitoring of the user vs no monitoring.*
In a way, all applications that we discuss here monitor the user; yet, there is a difference between pure monitoring as the main objective vs monitoring as a starting point for further actions. Further, it can be distinguished between constant time-continuous monitoring and recognition/detection of a user state at one point in time. For example, (public) surveillance is an ethically sensitive use case where monitoring is the main objective; see below in section 3.1.

[±**immediate**]**:** *System reacts (immediately/delayed) during the interaction with users vs no system reaction, or delayed reaction after an actual interaction.*
When the system reacts in a [+**immediate**] fashion, requirements on performance and criticality are higher; of course, this depends as well on the type of reaction – a simple ‘emotional back-channelling’ is harmless. While [±**automatic**] is about system decisions and actions in general – not necessarily in form of reaction to the user or affected person, [±**immediate**] pertains to the user-system interaction.

[±**mirroring**]**:** *Users receive feedback about what was detected/recognised by the system (e. g., which emotional expression they displayed) vs system does not give any explicit feedback.*

This aspect does not only play a role for the functionality of certain applications, e. g., therapy, but is also relevant to **transparency** in general.

[±**interacting**]**:** *System engages in an interaction with users, usually in form of a dialogue vs no interaction between system and user.*

This is a special type of [+**immediate**] reaction, with the same requirements. [+**interacting**] goes beyond a simple reaction or feedback by establishing a multi-turn interaction between system and user.

[±**affective**]**:** *System reacts displaying some 'affective' state vs system does behave 'neutral'.*

'Affective' here means showing emotions or 'emotion-related' states such as frustration, boredom, politeness, interest, care, to mention a few. This is a special type of [+**immediate**] reaction, with the same requirements. Research on conversational agents has shown that affective system responses make interactions more natural and therefore a more engaging and enjoyable experience (Diederich, Janssen-Müller, Brendel, & Morana, 2019). However, this can cause a tension between human-like, enjoyable interaction and the 'illusion of humanness', i. e., pretending emotional intelligence that a technical artefact does not possess. **Transparency** needs to be ensured to avoid deceiving the user.

2.4. Users / test set and subject areas

Fig. 1 **U** represents users and phenomena addressed. In the research flow, they stand for independent test samples, in the application flow, for real users or affected persons.

[±**typical**]**:** *Users are typical – often 'normal' adults, vs users are atypical, e. g., are members of minorities or stigmatised groups, or speakers with speech/language or other types of pathologies.*

[**-Typical**] groups are often more vulnerable, thus requirements on anonymity and **autonomy** are higher. This is on the one hand simply due to the fact that de-anonymisation for smaller groups with specific characteristics can be easier, because of the smaller search space. On the other hand, such groups are often more vulnerable per se, e. g., children in general and especially those with handicaps. Thus, ethical requirements can be lower when only typical users are addressed but have to be higher for atypical ones.

[±**generic**]**:** *The data produced by our subjects are 'generic' such as read, prompted lists vs the data are more specific, revealing more personal information.*

When the subject area, i. e., the phenomena we are interested in and that we record and process, are of a more general kind such as reading 'the Northwind and the sun', then these data reveal not much about the speaker. It is different when we, e. g., record free conversation where the speakers reveal personal information that can make them more easily identifiable, or when the application is aimed at eliciting them. The more generic the subject area is, by tendency, the less critical is the application for privacy.[9]

[9] Different constellations of types of users and types of data recorded, with more generic ones being less prone

3. Typology of Applications

Fig. 1 **T** shows a typology of applications.[10] Based on the degree of interaction, we distinguish four types: (1) **assessment**: no interaction between user and application, which just extracts information from the user's speech (or language) – the user might be informed about the assessment's results; (2) **reaction**: the system reacts to user actions or extracted information; (3) **interaction**: application and user interact; (4) **training**: a repeated interaction with a specific aim; the user is aware of this aim.

Note that there can be a distinction between user and affected person ('affectee'), especially for certain applications within **assessment** and **reaction**. Further, affectees might be unaware of being monitored compared to knowingly using an application.

In this section, we review existing applications and discuss the most crucial aspects from Fig. 1 for each application type. Fig. 2 depicts these application categories within our typology. In contrast to the review in Garcia-Garcia, Penichet, and Lozano (2017), which gives a high-level overview on available products for emotion recognition, we do not focus on individual companies or products, but attempt to categorise existing applications and highlight critical ethical and related implementation aspects for each category, according to our proposed taxonomies.

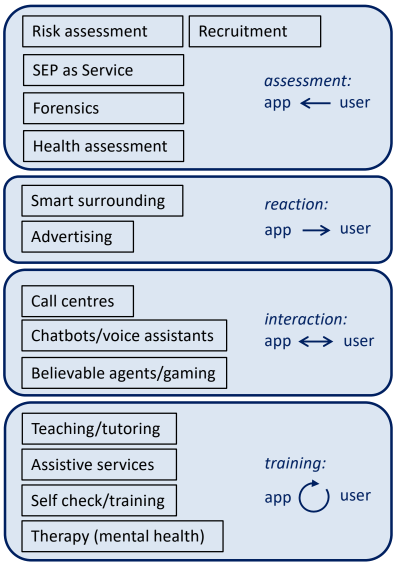

Figure 2. Detailed Typology of Applications.

3.1. Assessment

Risk assessment: This category spans a large variety of use cases, from monitoring individuals for a specific reason (e. g., credit risk **assessment**) to automated audio surveillance (Crocco, Cristani, Trucco, & Murino, 2016). Monitoring employees'

to violations of privacy, are discussed in Batliner et al. (2020).

[10] We use the term *application* in its broader meaning that is not limited to market-ready 'apps' like a smartphone app but also includes *use cases* in which CP systems are employed.

moods/feelings is claimed to improve decision making and performance, e. g., reducing trading risk in financial markets (Mayew & Venkatachalam, 2013; Whelan, McDuff, Gleasure, & Vom Brocke, 2018). Similar **assessment** applications are credit risk **assessment** and fraud detection (Gopinathan, Chaitanya, Kumar, & Rangarajan, 2015). Note that existing systems for lie detection have been criticised heavily (Eriksson & Lacerda, 2007; Kreiman & Sidtis, 2011). Automated surveillance of whole populations (e. g., as it is done in China) is the most serious type of **risk assessment**, leading to unethical consequences. Surveillance is mostly done by means of video data, but audio plays an important role as well, e. g., in auditory scene analysis (Crocco et al., 2016). Not all audio surveillance is unethical per se; there are also use cases for civil safety (Clavel, Vasilescu, Devillers, Richard, & Ehrette, 2008). Hence, critical ethical evaluation is always necessary (cf. Sec. 5). In all these **risk assessment** applications, the user of the system is not the affectee, and affectees are likely unaware of the **assessment**. Similar to recruitment, they are [+**critical**] and can be [+**intrusive**], and therefore need to be [+**interpretable**] and [+**accountable**] because of potential serious consequences.

Recruitment: In **recruitment**, voice analysis is said to provide unbiased personality tests of candidates; yet, cf. (Raghavan, Barocas, Kleinberg, & Levy, 2020). Regarding data representations, such applications need to be based on [+**representative**] and unbiased data. Ethically, they are [+**critical**] and can be [+**intrusive**], especially economically. We advocate that these applications have to be [+**interpretable**] and [+**accountable**] because they can cause serious consequences.

SEP as a Service: Existing application programming interfaces (APIs) for SEP mainly use emotion categories. However, as discussed in Sec. 2.1, representations depend on the task and the use of clear and strong emotion representations in real-world applications is questionable. Underlying datasets and methods are mostly undisclosed, leading to [–**interpretable**] and potentially [–**representative**] applications. Other ethically critical aspects are veneer of accuracy (strong claims about performance without transparent evidence), and cross-cultural awareness; claims about language-independence are questionable because, despite universal characteristics of emotions, cultural differences exist (Elfenbein & Ambady, 2003; Neiberg, Laukka, & Elfenbein, 2011)).

Health assessment: Applications that analyse a patient's speech with respect to specific health conditions are meant to help the treating physicians in detecting illnesses in early stages and monitoring the disease's progress, as well as to accelerate clinical trials. Such services exist, e. g., for Parkinson's disease (Klumpp et al., 2017) and Depression (Mdhaffar et al., 2019), among others; see Cummins, Baird, and Schuller (2018); Latif, Qadir, Qayyum, Usama, and Younis (2021) for an overview on speech analysis in healthcare. Crucial in health assessment is that the resulting predictions should not replace a diagnosis by a physician, and should thus be [–**automatic**]. They should further be [+**explainable**] and [+**interpretable**] to provide enough information for assisting the medical decision process, and respect the **autonomy** of the patient.

Forensics: Forensic evaluation of audio evidence like telephone calls, speaker identification, and verification can contribute to the conviction of criminals. The detection of single speaker traits can assist in this, e. g., of gender and age (Bahari & Van Hamme, 2011). These services need to be [+**interpretable**] and [+**explainable**] such that their results can be used in court (Hughes, Clermont, & Harrison, 2020). Especially in the case of speaker identification, they are [+**critical**] since wrong results may have serious consequences for the respective individuals. Thus, they have to be [+**performance**].

3.2. Reaction

Smart surrounding: Always-on, always-listening devices entered our lives recently (e. g., smart speakers, wearables) and can potentially function as emotion-aware environment, adapting to the user's mood. Two prevailing areas for such **smart surroundings** that we identified are smart home and automotive applications (Eyben et al., 2010). This application category is [–**monitoring**], as tracking emotions or other traits is not the main goal but a means to increase user satisfaction or to react to **critical** situations. Such systems need to be [+**immediate**] and [+**automatic**] and could be [+**personalising**]. These types of applications react to the users' state (with a certain action) but are usually **[-affective]** themselves. Ethically relevant are most notably data **privacy** (cf. Gray (2016)) and **transparency**. Applications should be [+**mirroring**] to inform the user why an action was taken. Regarding **privacy**, [+**on-device**] technologies are favourable, however, most products are cloud-based. Problematic uses of an emotion-aware surrounding arise if information is not used for the users' but for the company's benefit, which brings us to the next application category.

Advertising: Emotions play a crucial role in **advertising** (Holbrook & O'Shaughnessy, 1984; Poels & Dewitte, 2006). One actively explored strategy is dynamic placement of ads and adapting content based on a person's emotions (Chung, Patwa, & Markov, 2012). These endeavours are mainly based on facial expressions, but there is potential to take vocal reactions into account, too. One example is a patent by Amazon suggesting that advertisements can be presented based on mood and/or health information obtained from vocal interactions (Jin & Wang, 2018). This application is [+**monitoring**] and [+**automatic**]. We view emotion-aware **advertising** as [+**intrusive**] by driving customers' decisions. It is most likely [–**accountable**] which is ethically problematic. Note that advertising ethics is a whole research field on its own which we cannot possibly cover here, cf. Drumwright and Murphy (2009).

3.3. Interaction

Call centres: Using speech emotion recognition to monitor customer satisfaction in call centres is an application envisioned since the early days of SEP (Petrushin, 1999) and available as a commercialised service nowadays. SEP is used to monitor both customers and call centre agents, gaining insights about customer satisfaction and service quality. This application is [+**monitoring**], [+**interacting**], and [+**mirroring**] if agents get feedback. For customers, it is [–**intrusive**]; however, **transparency** needs to be ensured. In contrast, it can be [+**intrusive**] for employees, because constant monitoring puts pressure on them (Burkhardt, Huber, & Batliner, 2007). In Burkhardt, Engelbrecht, van Ballegooy, Polzehl, and Stegmann (2009), some strategies to deal with automatically detected user emotions in an interactive voice response system are discussed.

Chatbots/voice assistants: Regarding the proposed taxonomies, we treat **chatbots** and **voice assistants** as one application category. Key aspects in system design are that they are [–**monitoring**] (not the main goal), [+**immediate**], [+**automatic**] and [+**interacting**]. They can be [+**personalising**] and [+**affective**]. For instance, Amazon's Alexa can speak with different emotions (Aggarwal, Cotescu, Prateek, Lorenzo-Trueba, & Barra-Chicote, 2020; Gao, 2019). Ethically, we identified [+**representative**], **pri-**

vacy, and **transparency** as pivotal aspects. As we discussed representativity with respect to recognition of speaker states and traits so far, we highlight the importance for generation here, looking at systems that learn from user interactions (learning software). A negative example is the chatbot Tay, which started tweeting offensive and racist sentences after a few hours of learning from users (Wolf et al., 2017). This experiment has shown that automatically learnt generation can be strongly **biased**, with problematic outcomes.

Transparency and informational self-determination are crucial for these widespread consumer products. We advocate for opt-in mechanisms whenever unexpected additional information is processed, to reduce the "risk of unintended information disclosure" (Kröger, Lutz, & Raschke, 2019). In general, these applications are [–**intrusive**]. Exceptions are for example social chatbots like Xiaoice with which users converse for many hours (L. Zhou, Gao, Li, & Shum, 2020); they raise concerns of over-investment of time and addiction. A survey on social characteristics in chatbots can be found in Chaves and Gerosa (2020).

Believable agents/gaming: An important quality of believable characters and thus, believable agent, is "appropriately timed and clearly expressed emotion" (Bates, 1994). Hence, **believable agents** are [+**affective**] as well as [+**interacting**] and [+**personalising**]. We view them as umbrella term for many applications, e. g., social and health care **chatbots**, game characters, and motivating tutoring systems. Hence, ethical principles vary based on the application. However, the concept of believable agents in general is [–**intrusive**] and [–**critical**]. In gaming, SEP is used to create believable characters, and to alter the difficulty or character behaviour based on the player's emotion or stress level (Jones & Sutherland, 2008; Lobel et al., 2016).

3.4. Training

Teaching/tutoring: An understanding tutor can be viewed as a form of believable agent and thus classified accordingly. It needs to be [+**affective**] to motivate students and [+**mirroring**] to give feedback and act accordingly. Language learning is one specific use case with many available applications in this category, see for example (Randall, 2019) for a survey on language learning with robots.

Assistive services: Emotion-aware assistive services are mostly found in robotics (Ayari, Abdelkawy, Chibani, & Amirat, 2017). Nursing robots for elderly are one example. Another exemplary group of applications targets autism patients who have difficulties to communicate emotionally (Baron-Cohen, Golan, & Ashwin, 2009). Assistive services are [+**mirroring**] and potentially [+**personalising**] and [+**affective**]. Applications that assist the user in understanding others' emotions are [+**critical**].

Self check/training: This category comprises applications that give feedback to the user about their emotional expressions, see the 'emotional mirror' (Batliner et al., 2006), or social communication skills (Bosman, Bosse, & Formolo, 2019). It is [+**mirroring**] and [+**critical**] because the targeted paralinguistic phenomenon is the central aspect and it needs to be [+**interpretable**] and [+**explainable**] to give the user automatic feedback about speech features. This application is [–**intrusive**]. Training data can be acted and, e. g., represent strong, prototypical emotions on purpose.

Therapy in mental health: For mental disorders which affect a patient's emotional state or how a person perceives emotions, SEP provides potential for self-care applications, remote access to care, and in assisting diagnostics. One example is the

PRIORI project, addressing patients with bipolar disorder. Speech analysis on phone conversations (data in the wild) is used to detect mood state changes by bridging from recognised arousal/valence levels to mood states (Matton, McInnis, & Provost, 2019; Provost, Mcinnis, Gideon, Matton, & Khorram, 2020). Gaffney, Mansell, and Tai (2019) provide a review on conversational agents for mental health. System designs of applications in this space depend largely on the specific use case. Ethically, these applications are [+**critical**] because the outcomes can have a severe impact on people's well-being. They have to be [+**interpretable**], [+**explainable**], and [+**accountable**], especially for computer-aided diagnostics. Applications directly aiming at improving the patient's health state are [+**intrusive**], as opposed to pure monitoring tools that assist practitioners.

4. Interaction between taxonomies

Our taxonomies should not be handled in isolation as they can and often do influence each other. Changing one feature of a system may have an effect on aspects of several taxonomies, and the position on the scale of one aspect, i. e., its tendency towards + or –, can constrain the possible movements on the scale of another one. Some of these interactions are automatic and apply to applications in general but in most cases they also depend on other factors, like the specific task or target group, and should be reflected in the light of the actual situation.

In Tables 1 and 2, we list relations that are likely to exist in an application. They are not necessarily true in all situations but there are at least stronger tendencies towards them. We present three types of relations: (i) *causal* relations in which one criterion directly influences another one, denoted with '$\rightarrow$'; (ii) *correlations* in which two values are likely to appear in certain combinations together, denoted with '$\approx$'; and (iii) *increase of relevance* which means that the value of one taxonomy makes an ethical cornerstone more important to be considered, denoted with '$\Rightarrow$'. For each relation, the aspect that likely has the higher influence over the other one stands on the left side of the symbol, the other term on the right. For (i) $\rightarrow$ and (iii) $\Rightarrow$, the left term is the cause or trigger in the relation. For correlations in (ii) $\approx$, the relation is more conditioned on the left than on the right aspect, i. e., the right value is more likely to occur together with the left one than vice versa, but it is not a causal dependency. In each interaction in the table, both the left and the right side can consist of several terms and are marked with a tendency towards + or –. The only exception to the latter are the ethical cornerstones in (iii) $\Rightarrow$, because the relation increases the relevance of the cornerstone without moving towards one of the endpoints of the scale.

Now we want to give a few examples to demonstrate the idea behind the listed interactions and their limitations. For instance, the second entry in Table 1, **E** $\longrightarrow$ **E**: [+**explainable**]/[+**interpretable**] $\rightarrow$ [–**biased**] means that making a model explainable or interpretable can help to detect and thus handle biases during development. Of course, explainability and interpretability do not guarantee a system without any biases – it might be impossible to detect all existing biases, and in the end, it depends on how the developer reacts on them. Understanding the model and the representations it is based on can even result in more bias to compensate unfairness (cf. Sec. 2.2). In general, however, biases are unwanted, and if they can be detected by explainable or interpretable models, they will be most likely tried to be reduced.

The interaction **S** $\longrightarrow$ **S:** [+**interacting**] $\rightarrow$ [+**immediate**] holds when the main actions of a system happen during the interaction with the user. This is often the case

E* ⟶ *E
[+**performance**] → [+**biased**]: *With high performance, it is more likely that the data are more biased and less representative for the targeted population.*
[+**explainable**]/[+**interpretable**] → [–**biased**]: *Making the inner processes of a model more transparent can help to detect and handle unwanted distribution biases, resulting in less biased models.*
[–**explainable**]/[–**interpretable**] → [–**accountable**]: *Full accountability is difficult to achieve when it is not possible to understand why and how the system came to its predictions.*
[+**intrusive**] → [+**critical**]: *Intrusive applications are more likely to be critical since wrong decisions might influence the user in an undesired and unethical way.*
[+**critical**] ⇒ [**performance**]: *A high performance is crucial in critical applications as it is important to decrease the failure rate as much as possible.*
[+**critical**] ⇒ [**representative**]: *For critical applications, it is most important that the training data are representative for the targeted population.*
[+**critical**] ⇒ [**explainable**]/[**interpretable**]/[**accountable**]: *In critical applications, it is important to make the system's decision process transparent to be able to evaluate the predictions and their appropriateness.*
[+**intrusive**] ⇒ [**explainable**]/[**accountable**]: *Since intrusive applications can have severe impact on the user's state, it is more important to provide explanations.*
S* ⟶ *S
[+**online learning**] → [–**on-device**]: *Online learning tends to be performed on external servers with powerful hardware and profit from more, possibly user-independent data.*
[+**immediate**] → [+**automatic**]: *For a system reaction happening within an interaction, human intervention is barely possible, except for actions taken afterwards.*
[+**interacting**] → [+**immediate**]: *Interacting systems are more likely to react immediately instead of delayed.*
[+**personalising**] ≈ [+**online learning**]: *A personalised system is often connected to an online learning mechanism to increase the relevance of decisions for a specific user.*
[+**affective**] ≈ [+**immediate**]: *Affective systems often display emotions and the like as immediate reactions to the user.*

Table 1. Automatic interactions *within taxonomies*, denoted with ⟶. → denotes a causal relation that might exist necessarily or just with high probability, ≈ marks a correlation between taxonomies, and ⇒ stands for a triggered increase of relevance of an ethical cornerstone. Note that [+] and [–] do not always indicate the extremes but tendencies.

$\boldsymbol{S} \longrightarrow \boldsymbol{E}$

[–**on-device**] → [–**anonymous**]: *If user data is transferred to the cloud, the service needs to store personal information like the IP address to send results back to the user.*

[+**automatic**]/[+**immediate**] → [+**critical**]: *An application that reacts immediately and/or automatically is more likely to be critical since no human intervention is possible.*

[+**online learning**] → [–**representative**]/[+**biased**]: *In online learning, controlling the training data is difficult, resulting in a higher risk for learning biases and unrepresentative distributions*

[+**personalising**] → [–**representative**]: *A personalised application is tailored to specific speakers and cannot be representative for a whole population.*

[+**personalising**] → [–**anonymous**]: *Personalising systems are more likely to fail to provide sufficient anonymity and might need to actively incorporate anonymisation strategies.*

[+**affective**] → [+**intrusive**]: *A system that uses affective language can influence the emotional state of the user and thus be intrusive.*

[+**automatic**] ⇒ [**performance**]: *If a system acts in an automatic ways, an ethically reasonable performance is more important because no human can intervene in the process.*

[+**automatic**] ⇒ [**accountable**]: *If no human intervention is possible between prediction and reaction, solving the question of accountability becomes more important.*

$\boldsymbol{U} \longrightarrow \boldsymbol{E}$

[–**typical**] → [–**anonymous**]: *With decreasing size of a group sharing a specific attribute (e. g., a disease), the easier it is to identify individuals in corresponding representations*

[–**typical**] ⇒ [**representative**]/[**biased**]: *When dealing with minorities, unwanted biases and missing representativity are more critical and thus more important to be controlled.*

[–**generic**] ≈ [+**biased**]/[–**anonymous**]: *There is a higher risk for unwanted biases and hidden private information in the data if the experimental setting is less generic.*

$\boldsymbol{R} \longrightarrow \boldsymbol{E}$

[+**proximal**] → [+**representative**]: *Trivially, the closer the proxy is to the phenomenon, the more representative is the modelling.*

[–**real**] → [–**representative**]: *A system cannot really be representative if the data is not realistic.*

[+**speaker-dependent**] → [–**representative**]: *Trivially, a speaker-dependent modelling cannot account for representativity.*

[+**real**] ⇒ [**anonymous**]: *If real data is used, we have to care about anonymity of the subjects.*

[+**real**] ≈ [+**biased**]: *Since classes are often not distributed equally in real data, there might be biases towards certain classes in it.*

Table 2. Automatic interactions *between taxonomies*, denoted with $\longrightarrow$. → denotes a causal relation that might exist necessarily or just with high probability, ≈ marks a correlation between taxonomies, and ⇒ stands for a triggered increase of relevance of an ethical cornerstone. Note that [+] and [–] do not always indicate the extremes but tendencies.

for dialogue systems that mainly inform the user about something or process requests. An example in which this relation does not exist can be found in health assessment if a system tracks the user's health state over several conversations and transmits the prediction afterwards to the corresponding practitioner.

The correlation **R** $\longrightarrow$ **E:** [**+real**] $\approx$ [**+biased**] means that realistic data are often the basis for more **biased** systems because properties are not distributed equally in the real world, and thus in a realistic setting not in the model either. For instance, an emotion recognition model will most likely have some bias towards certain emotion classes or dimensions if it was trained on real (i. e., unbalanced) data. Real data do not necessarily contain such unequal distributions but it is more likely to find them there than in data from fully controlled experiments.

As a final example, **S** $\longrightarrow$ **E:** [**+automatic**] $\Rightarrow$ [**performance**] means that it becomes more relevant to carefully assess a system's performance if it reacts automatically. More specifically, if no human processes the model's prediction before the final action, i. e., nobody could intervene in a false reaction, it is important that the system has a performance that is reasonable from the ethical point of view. The performance threshold could be quite low if the application is ethically lightweight and thus [**–critical**]. The interaction does not mean that **performance** can be neglected if the system is not automatic, just that this property influences the importance of the cornerstone.

5. Red Lines and Ethics Washing

As far as we can see, the applications that we describe or are envisioning are not unethical per se which would mean that they should be banned from the very beginning, such as autonomous weapons; yet, types of applications such as **risk assessment** could be expanded to surveillance of public places in such a way that society should not only be aware but put up *red lines* against them. Such red lines are much debated in the public discourse and were addressed and aimed at in EU regulations but have been weakened by lobbyists resulting in *ethics washing*, i. e., 'performative' consideration to ethics (Bietti, 2020; Rességuier & Rodrigues, 2020; Wagner, 2018), along the well-known lines of oil companies and Big Tobacco (Proctor, 2008). In general, for the field of ethical AI, we see currently in both the literature and media a deeply-rooted intention by the research community (both academic and industry) to address the many aspects of the cornerstones mentioned above. However, implementing such change is currently much more superficial, with minimal change in the last years, and the rehashing of similar topics.

We consider that there appears to be a significant difference between corporate responsibility and genuine *ethical awareness* (Benjamins et al., 2019). In other words, as yet, it seems that ethics in areas of AI are instead a set of *buzz words* to build commercial trust, show intention, or address a specific responsibility rather than low-level sincere concern. This becomes evident in cases of conflicts between employees and companies[11] (Ebell et al., 2021).

Despite this, as we mention, there have been regulations at a higher governmental level in recent time that may encourage a faster change, and foster a more honest intention. For example, the European Commission presented their guidelines for Artificial

[11] https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/, retireved 05/03/2021.

Intelligence[12] – in which they suggest the explicit banning of *manipulative* AI systems which may implement aspects of social scoring, limiting biometric-based real-time remote identification, as well as "special transparency requirements on all emotion recognition and biometric categorisation systems".

6. Caveats and Limitations

From our current contribution, we would like to make the reader aware of a few limitations and caveats that we have come across throughout our evaluation. Of most prominence, we cannot outline an entirely exhaustive overview of speech-based applications, given the vast number of use-cases which are currently being developed; neither is it possible to give an exhaustive account of critical concepts such as XAI – for example, in Adadi and Berrada (2018); Arrieta et al. (2020), the authors list hundreds of publications which refer to XAI. We attempt to highlight aspects that are most prominent for CP, but cannot concretely suggest that we cover all aspects. In this way, we should be clear that this contribution does not rank the value of single ethical aspects – with a points system or similar – as there are far too many diverse variables that might constitute the value from addressing each of the combined cornerstones. Thus, we chose the approach of cornerstones itself, considering each to have an individual merit, which, when combined as a whole, covers a vast majority of the current ethical criteria related to CP. Furthermore, we did not try to compare our taxonomies against realistic expectation; e. g., patents might prevent full accountability by a developer. Moreover, a complete account of data and models in an *ethics card* might be unrealistic in certain scenarios, due to time/financial constraints, and the entire effort needed. This might change in the long run when ethics regulations in our field will – hopefully – have been discussed, established, and decided upon in more detail, so they approach at least the same level of standardisation as is the case for medical and pharmacological research and applications (Mittelstadt, 2019).

7. Concluding Remarks

For applications within Computational Paralinguistics, we presented five different taxonomies covering data representation, system design, cornerstones of ethical aspects, a typology based on the degree of interaction, and users and subject areas; these are evaluated with regard to the principles of **principalism**. The taxonomies are intended to help assessing applications as for their specific requirements on ethics; they are not a definite set – yet, they cover the relevant aspects and intend to be a basis for more elaborated accounts. Moreover, they can be employed as a sort of checklist for people creating applications, using such applications, or reading about them.

8. Acknowledgments

We acknowledge funding from the EU's Horizon 2020 programme, grant agreement No. 826506 (sustAGE), and the DFG's Reinhart Koselleck project No. 442218748

[12] Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act); https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence [Accessed May 07, 2021]

(AUDI0NOMOUS).

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138-52160.

Aggarwal, V., Cotescu, M., Prateek, N., Lorenzo-Trueba, J., & Barra-Chicote, R. (2020). Using VAEs and Normalizing Flows for One-Shot Text-To-Speech Synthesis of Expressive Speech. In *Proc. of ICASSP* (p. 6179-6183).

Arrieta, A. B., Rodríguez, N. D., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115.

Ayari, N., Abdelkawy, H., Chibani, A., & Amirat, Y. Y. (2017). Towards Semantic Multimodal Emotion Recognition for Enhancing Assistive Services in Ubiquitous Robotics. In *Proc. of the AAAI 2017 Fall Symposium Series* (p. 2-9). Arlington, United States.

Bahari, M. H., & Van Hamme, H. (2011). Speaker age estimation and gender detection based on supervised non-negative matrix factorization. In *2011 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS)* (p. 1-6).

Barnett, I., & Torous, J. B. (2019). Ethics, Transparency, and Public Health at the Intersection of Innovation and Facebook's Suicide Prevention Efforts. *Annals of Internal Medicine*, *170*, 565-566.

Baron-Cohen, S., Golan, O., & Ashwin, E. (2009). Can emotion recognition be taught to children with autism spectrum conditions? *Philosophical Transactions of the Royal Society B*, *364*, 3567–3574.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, *20*, 1-68.

Bates, J. (1994). The role of emotion in believable agents. *Communications of the ACM*, *37*, 122–125.

Batliner, A., Burkhardt, F., van Ballegooy, M., & Nöth, E. (2006). A Taxonomy of Applications that Utilize Emotional Awareness. In *Proceedings of is-ltc 2006* (pp. 246–250). Ljubliana.

Batliner, A., Hantke, S., & Schuller, B. W. (2020). Ethics and good practice in computational paralinguistics. *IEEE Transactions on Affective Computing*.

Batliner, A., & Möbius, B. (2020). Prosody in Automatic Speech Processing. In C. Gussenhoven & A. Chen (Eds.), *The oxford handbook of language prosody* (p. 633-645). Oxford, UK: Oxford University Press.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., ... Amir, N. (2011). Whodunnit – Searching for the Most Important Feature Types Signalling Emotional User States in Speech. *Computer Speech and Language*, *25*, 4–28.

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, *6*, 587-604.

Benjamins, R., Barbado, A., & Sierra, D. (2019). *Responsible AI by Design in Practice.* Retrieved from http://arxiv.org/abs/1909.12838

Bietti, E. (2020). From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (p. 210-219). Barcelona, Spain.

Bortolus, A. (2008). Error Cascades in the Biological Sciences: The Unwanted Consequences of Using Bad Taxonomy in Ecology. *AMBIO A Journal of the Human Environment*, *37*, 114-118.

Bosman, K., Bosse, T., & Formolo, D. (2019). Virtual agents for professional social skills training: An overview of the state-of-the-art. In P. Cortez, L. Magalhães, P. Branco, C. F. Portela, &

T. Adão (Eds.), *Intelligent Technologies for Interactive Entertainment* (pp. 75–84). Cham: Springer International Publishing.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 77–91). New York, NY, USA. Retrieved from `http://proceedings.mlr.press/v81/buolamwini18a.html`

Burkhardt, F., Engelbrecht, K. P., van Ballegooy, M., Polzehl, T., & Stegmann, J. (2009). Emotion Detection in Dialog Systems - Usecases, Strategies and Challenges. In *Proceedings of Affective Computing and Intelligent Interaction (ACII), Amsterdam, The Netherlands.*

Burkhardt, F., Huber, R., & Batliner, A. (2007). Application of Speaker Classification in Human Machine Dialog Systems. In C. Müller (Ed.), *Speaker Classification I Fundamentals, Features, and Methods* (pp. 174–179). Berlin-Heidelberg: Springer.

C., W. T. (1926). Taxonomy in Biology. *Nature*, *118*, 901-902.

Chancellor, S., Birnbaum, M., Caine, E., Silenzio, V., & Choudhury, M. D. (2019). A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. In *Proceedings of the conference on fairness, accountability, and transparency* (p. 7988). Atlanta, GA, USA.

Chaves, A. P., & Gerosa, M. A. (2020). How should my chatbot interact? a survey on social characteristics in humanchatbot interaction design. *International Journal of HumanComputer Interaction*, 1-30.

Chung, W. J., Patwa, P., & Markov, M. M. (2012, June 7). *Targeting advertisements based on emotion.* US Patent App. 12/958,775.

Clavel, C., Vasilescu, I., Devillers, L., Richard, G., & Ehrette, T. (2008). Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, *50*(6), 487-503.

Cowie, R. (2012). The good our field can hope to do, the harm it should avoid. *IEEE Transactions on Affective Computing*, *3*, 410-423.

Crawford, K., Roel Dobbe, T. D., Green, G. F. B., Kaziunas, E., Kak, A., Mathur, V., ... Whittaker, M. (2019). *AI Now 2019 Report.* https://ainowinstitute.org/reports.html. New York: AI Now Institute.

Crocco, M., Cristani, M., Trucco, A., & Murino, V. (2016). Audio surveillance: A systematic review. *ACM Computing Surveys (CSUR)*, *48*(4), 1–46.

Cummins, N., Baird, A., & Schuller, B. W. (2018). Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*, *151*, 41-54. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1046202317303717` (Health Informatics and Translational Data Analytics)

Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI-17* (pp. 4691–4697). Melbourne.

Dellaert, F., Polzin, T., & Waibel, A. (1996). Recognizing emotion in speech. In *Proc. of ICSLP* (p. 1970-1973). Philadelphia.

Diederich, S., Janssen-Müller, M., Brendel, A. B., & Morana, S. (2019). Emulating Empathetic Behavior in Online Service Encounters with Sentiment-Adaptive Responses: Insights from an Experiment with a Conversational Agent. In *Proc. of ICIS 2019 – Smart Service Systems and Service Science.* Munich, Germany.

Dobbe, R., Dean, S., Gilbert, T., & Kohli, N. (2018). A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics. *CoRR*, *abs/1807.00553*. Retrieved from `http://arxiv.org/abs/1807.00553`

Doran, D., Schulz, S., & Besold, T. R. (2017). *What Does Explainable AI Really Mean? A New Conceptualization of Perspectives.* Retrieved from `http://arxiv.org/abs/1710.00794`

Döring, S., Goldie, P., & McGuinness, S. (2011). Principalism: A Method for the Ethics of Emotion-Oriented Machines. In R. Cowie, C. Pelachaud, & P. Petta (Eds.), *Emotion-Oriented Systems: The Humaine Handbook* (pp. 713–724). Berlin, Heidelberg: Springer.

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning.* Retrieved from `https://arxiv.org/abs/1702.08608`

Drumwright, M. E., & Murphy, P. E. (2009). The current state of advertising ethics: Industry and academic perspectives. *Journal of Advertising*, *38*, 83–108.

Dumont, J., Giergiczny, M., & Hess, S. (2015). Individual level models vs. sample level models: contrasts and mutual benefits. *Transportmetrica A: Transport Science*, *11*, 465-483.

Ebell, C., Baeza-Yates, R., Benjamins, R., Cai, H., Coeckelbergh, M., Duarte, T., ... Thais, S. (2021). Towards intellectual freedom in an AI Ethics Global Community. *AI and Ethics*. Retrieved from https://doi.org/10.1007%2Fs43681-021-00052-5

Elfenbein, H. A., & Ambady, N. (2003). Universals and Cultural Differences in Recognizing Emotions. *Current Directions in Psychological Science*, *12*, 159-164.

Eriksson, A., & Lacerda, F. (2007). Charlatanry in forensic speech science: A problem to be taken seriously. *Journal of Speech, Language and the Law*, *14*, 169–193.

Eyben, F., Weninger, F., Groß, F., & Schuller, B. (2013). Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In *Proc. ACM Multimedia* (pp. 835–838). Barcelona, Spain.

Eyben, F., Wöllmer, M., Poitschke, T., Schuller, B., Blaschke, C., Färber, B., & Nguyen-Thien, N. (2010). Emotion on the roadnecessity, acceptance, and feasibility of affective computing in the car. *Advances in human-computer interaction*, *2010*.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, Á., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI.* Retrieved from https://dash.harvard.edu/handle/1/42160420 (Berkman Klein Center for Internet & Society)

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*, 689–707.

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

Friedman, B. (1996). Value-Sensitive Design. *Interactions*, *3*, 16-23.

Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. *Transactions on Information Systems*, *14*, 330-347.

Gaffney, H., Mansell, W., & Tai, S. (2019). Conversational agents in the treatment of mental health problems: Mixed-method systematic review. *JMIR Ment Health*, *6*, e14166.

Gao, C. (2019). *Use New Alexa Emotions and Speaking Styles to Create a More Natural and Intuitive Voice Experience.* Retrieved from https://developer.amazon.com/en-US/blogs/alexa/alexa-skills-kit/2019/11/new-alexa-emotions-and-speaking-styles

Garcia-Garcia, J. M., Penichet, V., & Lozano, M. (2017). Emotion detection: a technology review. In *Proceedings of the XVIII International Conference on Human Computer Interaction (Interacción 17).* (8 pages)

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H. M., Daumé III, H., & Crawford, K. (2018). *Datasheets for Datasets.* Retrieved from http://arxiv.org/abs/1803.09010

Ghahari, H., Carpintero, D. L., Ostovan, H., Kolarov, J., Collingwood, C., Finlayson, T., & Fischer, M. (2010). Ethics and accuracy in scientific researches with emphasize on taxonomic works. *Linzer Biologische Beiträge*, *42*, 671-694.

Gopinathan, K. R., Chaitanya, J., Kumar, S. R., & Rangarajan, S. (2015, May 21). *Credit risk decision management system and method using voice analytics.* US Patent App. 14/549,505.

Gray, S. (2016). Always on: privacy implications of microphone-enabled devices. In *Future of privacy forum.* Washington, DC: FPF.

Holbrook, M. B., & O'Shaughnessy, J. (1984). The role of emotion in advertising. *Psychology & Marketing*, *1*, 45–64.

Holstein, K., Vaughan, J. W., Daumé III, H., Dudík, M., & Wallach, H. M. (2018). *Improving fairness in machine learning systems: What do industry practitioners need?* Retrieved from http://arxiv.org/abs/1812.05239

Hughes, V., Clermont, F., & Harrison, P. (2020). Correlating Cepstra with Formant Frequencies: Implications for Phonetically-Informed Forensic Voice Comparison. In *Proc. of Interspeech* (pp. 1858–1862).

Jin, H., & Wang, S. (2018, October 9). *Voice-based determination of physical and emotional characteristics of users.* U.S. Patent No. 10,096,319.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, *1*, 389–399.

Johnson, D. G., & Wetmore, J. M. (2008). STS and Ethics: Implications for Engineering Ethics. In E. J. Hackett, O. Amsterdamska, M. Lynch, & J. Wajcman (Eds.), *The Handbook of Science and Technology Studies, 3rd ed.* (p. 567-581). Cambridge, Massachusetts and London, England: MIT Press.

Jones, C., & Sutherland, J. (2008). Acoustic emotion recognition for affective computer gaming. In *Affect and emotion in human-computer interaction* (pp. 209–219). Springer.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Vigas, F. B., & Sayres, R. (2018). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In J. G. Dy & A. Krause (Eds.), *Icml* (Vol. 80, p. 2673-2682). PMLR.

Klumpp, P., Janu, T., Arias-Vergara, T., Vsquez-Correa, J., Orozco-Arroyave, J. R., & Nth, E. (2017). Apkinson A Mobile Monitoring Solution for Parkinsons Disease. In *Proc. of Interspeech* (pp. 1839–1843).

Kreiman, J., & Sidtis, D. (2011). *Foundations of Voice Studies - An Interdisciplinary Approach to Voice Production and Perception.* Wiley & Sons.

Kröger, J. L., Lutz, O. H.-M., & Raschke, P. (2019). Privacy Implications of Voice and Speech Analysis–Information Disclosure by Inference. In *IFIP International Summer School on Privacy and Identity Management* (pp. 242–258).

Krug, A., & Stober, S. (2018). Introspection for convolutional automatic speech recognition. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 187–199).

Latif, S., Qadir, J., Qayyum, A., Usama, M., & Younis, S. (2021). Speech Technology for Healthcare: Opportunities, Challenges, and State of the Art. *IEEE Reviews in Biomedical Engineering*, *14*, 342-356.

Leroy, D., Coucke, A., Lavril, T., Gisselbrecht, T., & Dureau, J. (2019). Federated learning for keyword spotting. In *Proc. of ICASSP* (pp. 6341–6345).

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector.* The Alan Turing Institut. Retrieved from `https://doi.org/10.5281/zenodo.3240529`

Lipton, Z. C. (2016). The Mythos of Model Interpretability. *CoRR*, *abs/1606.03490*. Retrieved from `http://arxiv.org/abs/1606.03490`

Lobel, A., Gotsis, M., Reynolds, E., Annetta, M., Engels, R. C., & Granic, I. (2016). Designing and utilizing biofeedback games for emotion regulation: The case of nevermind. In *Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 1945–1951).

Lo Piano, S. (2020). Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanit Soc Sci Commun*, *7*. (7 pages)

Matton, K., McInnis, M. G., & Provost, E. M. (2019). Into the Wild: Transitioning from Recognizing Mood in Clinical Interactions to Personal Conversations for Individuals with Bipolar Disorder. In *Proc. of Interspeech* (p. 1438-1442). Graz.

Mayew, W. J., & Venkatachalam, M. (2013). Speech analysis in financial markets. *Foundations and Trends® in Accounting*, *7*, 73–130.

Mdhaffar, A., Cherif, F., Kessentini, Y., Maalej, M., Thabet, J. B., Maalej, M., ... Freisleben, B. (2019). Dl4ded: Deep learning for depressive episode detection on mobile devices. In J. Pagán, M. Mokhtari, H. Aloulou, B. Abdulrazak, & M. F. Cabrera (Eds.), *How ai impacts urban living and public health* (pp. 109–121). Cham: Springer International Publishing.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... Gebru, T. (2018). *Model Cards for Model Reporting.* Retrieved from `http://arxiv.org/abs/`

1810.03993

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nat Mach Intell*, *1*, 501507.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency* (p. 279288). Atlanta, GA.

Molnar, C., Casalicchio, G., & Bischl, B. (2020). *Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges.* Retrieved from https://arxiv.org/abs/2010.09337

Neiberg, D., Laukka, P., & Elfenbein, H. A. (2011). Intra-, inter-, and cross-cultural classification of vocal affect. In *Proc. of Interspeech* (p. 1581-1584). Florence.

Newbold, A., Warren, F. C., Taylor, R. S., Hulme, C., Burnett, S., Aas, B., . . . Watkins, E. R. (2020). Promotion of mental health in young adults via mobile phone app: Study protocol of the ECoWeB (emotional competence for well-being in Young adults) cohort multiple randomised trials. *BMC Psychiatry*, *20*.

Petrushin, V. (1999). Emotion in speech: Recognition and application to call centers. In *Proceedings of artificial neural networks in engineering* (Vol. 710, p. 22).

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological applications*, *19*, 181–197.

Picard, R. (1997). *Affective Computing.* Cambridge, MA: MIT Press.

Poels, K., & Dewitte, S. (2006). How to capture the heart? reviewing 20 years of emotion measurement in advertising. *Journal of Advertising Research*, *46*, 18–37.

Proctor, R. N. (2008). Agnotology: A Missing Term to Describe the Cultural Production of Ignorance (and Its Study). In R. N. Proctor & L. Schiebinger (Eds.), *Agnotology: The Making and Unmaking of Ignorance* (p. 1-36). Stanford, CA: Stanford University Press.

Provost, E. M., Mcinnis, M., Gideon, J. H., Matton, K. A., & Khorram, S. (2020, March 5). *Automatic speech-based longitudinal emotion and mood recognition for mental health treatment.* US Patent App. 16/556,858.

Raghavan, M., Barocas, S., Kleinberg, J. M., & Levy, K. (2020). Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 469–481).

Randall, N. (2019). A Survey of Robot-Assisted Language Learning (RALL). *J. Hum.-Robot Interact.*, *9*. Retrieved from https://doi.org/10.1145/3345506 (36 pages)

Ravanelli, M., & Bengio, Y. (2019). *Interpretable Convolutional Filters with SincNet.* Retrieved from https://arxiv.org/abs/1811.09725

Resnick, B. (2019). *Yes, artificial intelligence can be racist.* Retrieved from https://www.vox.com/science-and-health/2019/1/23/18194717/alexandria-ocasio-cortez-ai-bias/

Rességuier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society*, *7*, 1-5.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier.* Retrieved from http://arxiv.org/abs/1602.04938

RichardBenjamins. (2021). A choices framework fortheresponsible use ofAI. *AI and Ethics*, *1*, 49-53.

Samyn, Y., & Massin, C. (2002). Taxonomists' Requiem? *Science*, *295*, 276-277.

Schröder, M. (2001). Emotional speech synthesis: a review. In *Proc. of Eurospeech* (p. 561-564). Aalborg.

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

Stahl, B. C., Timmermans, J., & Mittelstadt, B. D. (2016). The Ethics of Computing: A Survey of the Computing-Oriented Literature. *ACM Comput. Surv.*, *48*. (38 pages)

Stappen, L., Baird, A., Christ, L., Schumann, L., Sertolli, B., Messner, E.-M., . . . Schuller, B. W. (2021). *The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress.* Retrieved from https://arxiv.org/abs/2104.07123

Taxonomy. (1992). In *Encyclopedia Britannica: Micropedia* (Vol. 11, p. 586). Chicago. (15th edition)

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2020). Implementations in Machine Ethics: A Survey. *ACM Computing Surveys*, *53*, 138.

Tomashenko, N., Srivastava, B. M. L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., … Todisco, M. (2020). Introducing the VoicePrivacy initiative. In *Proc. of Interspeech* (p. 1693-1697). Shanghai, China.

Torous, J., & Roberts, L. W. (2017). The Ethical Use of Mobile Health Technology in Clinical Psychiatry. *J Nerv Ment Dis*, *205*, 4-8.

Wagner, B. (2018). Ethics as an Escape from Regulation: From ethics-washing to ethics shopping? In E. Bayamlioglu, M. Hildebrandt, I. Baraluic, & L. Janssen (Eds.), *Being profiled:cogitas ergo sum: 10 years of profiling the european citizen* (p. 84-89). Amsterdam, NL: Amsterdam University Press.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, *70*, 129–133.

Whelan, E., McDuff, D., Gleasure, R., & Vom Brocke, J. (2018). How emotion-sensing technology can reshape the workplace. *MIT Sloan Management Review*, *59*, 7–10.

Wolf, M. J., Miller, K., & Grodzinsky, F. S. (2017). Why we should have seen that coming: comments on microsoft's tay 'experiment,' and wider implications. *ACM SIGCAS Computers and Society*, *47*, 54–64.

World Health Organization (WHO). (1993). *The icd-10 classification of mental and behavioural disorders.* Geneva, Switzerland: World Health Organization.

Zhou, L., Gao, J., Li, D., & Shum, H.-Y. (2020). The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, *46*, 53–93.

Zhou, Y., & Danks, D. (2020). Different "Intelligibility" for Different Folks. In A. N. Markham, J. Powles, T. Walsh, & A. L. Washington (Eds.), *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY* (pp. 194–199).