Ethical considerations of the GermEval20 Task 1. IQ assessment with natural language processing: Forbidden research or gain of knowledge?

Dirk Johannßen
MIN Faculty,
Dept. of Computer Science
Universität Hamburg
& Nordakademie

Chris Biemann
MIN Faculty,
Dept. of Computer Science
Universität Hamburg
22527 Hamburg, Germany

David Scheffer
Faculty of Economics
NORDAKADEMIE
25337 Elmshorn, Germany

http://lt.informatik.uni-hamburg.de/
{biemann, johannssen}@informatik.uni-hamburg.de
david.scheffer@nordakademie.de

Abstract

The use of Intelligence Quotient (IQ) testing as a measure for intellectual ability is controversial. Even though IQ testing is considered to be among the most valid measures of psychology, findings and current research sparked a debate over racial or socioeconomic biases, as well as the label of 'pseudoscience' for many situations that involve IQ testing. The GermEval20 Task 1 asked researchers to investigate NLP approaches for approximating the ranking based on IQ and high school scores from so-called implicit motive texts. Quickly, a vivid discourse on whether the shared task should be viewed as unethical and forbidden was held within the NLP community. In this paper, we investigate ethical considerations and arguments against and in favor of such a task and come to the conclusion that this type of research should be conducted despite the undoubtedly associated ethical issues, as it can shed light on thus-far non-transparent methods and offers valuable gains of knowledge.

1 Introduction

The use of Intelligence Quotient (IQ) testing as a measure for intellectual ability is highly controversial. On the one hand, different IQ measures have been established by psychologists more than a century ago and are held to be among the most valid, stable, and reliable measures in the whole scientific field of psychology (Benson, 2003). Cognitive abilities are influential predictors for multiple criteria for professional success (Schmidt and Hunter, 1998; Kramer et al., 2009; Ones et al., 2017). On the other hand, recent studies have shown that many contucted IQ tests are fundamentally flawed as to introduce racial or socioeconomic biases (Turkheimer et al., 2003).

Even the term *intelligence* is ambiguous as well as the assumed impacts IQ testing has on aptitude diagnostics, as the title of the paper *'intelligence is what the intelligence test measures'* suggests (Maas et al., 2014), whilst most definitions at least agree that intelligence is always connected to successfully overcoming challenges of everyday life situations (Rost, 2009).

Since those aspects already are concerns, any work that involves automation by the use of NLP techniques and IQ testing information based on natural language texts understandably raises concerns. In case of the GermEval20 Task 1, the research has caused more than just concerns, but a so-called *shitstorm* on the social platform Twitter[1].

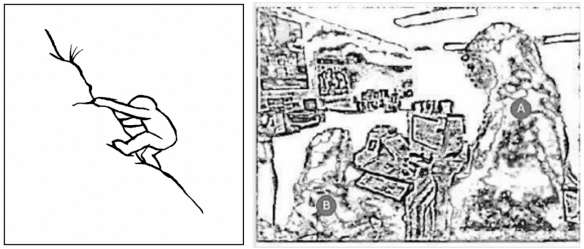

Figure 1: Images to be interpreted by participants utilized for the operant motive test (OMT) on the left and the motive index (MIX) on the right. The motives are the affiliation motive (A), the power motive (M), achievement (L), and freedom (F).A 0 stays for the zero / unassigned motive.

The *GermEval20 Task 1 on the Classification and Regression of Cognitive and Emotional Style from Text (Johannßen et al., 2020)*[2][3] was the stumbling block for a heated debate on this topic.

In short, the task is about researching the validity of so-called implicit motives and their potential to substitute controversial metrics utilized in

[1] https://www.twitter.com

[2] GermEval is a series of shared task evaluation campaigns that focus on Natural Language Processing (NLP) for the German language. The workshop is held as a joint Conference SwissText & KONVENS 2020 in Zürich.

[3] https://competitions.codalab.org/competitions/22006

aptitude diagnostics. Metrics like IQ tests, high school grades, or math tests are commonly used in aptitude diagnostics, but can quickly be utilized in inherently flawed settings. Research on implicit motives has shown, that they are indicators for long-time behavior and development, which could replace the other metrics (Scheffer, 2004). For those implicit motives, participants are asked to answer questions to ambiguous drawings (Scheffer and Kuhl, 2013). Those association tasks are less socioeconomical or racial biased and rather show intrinsic desires than task comprehension (Schultheiss, 2008, p. 439 ff.). However, as there is no clear and unambiguous rule system for labeling those implicit motives, they have yet to be understood better (Johannßen and Biemann, 2019). This is, what the shared task aimed: a better understanding of implicit motives and their role in aptitude diagnostics by the use of Natural Language Processing (NLP) methods.

The task contains two subtasks. For Subtask 1, participants are asked to reproduce a ranking of students based on different high school grades and IQ scores solemnly from implicit motive texts. For Subtask 2, participants are asked to classify each motive text into one of 30 classes as a combination of one of five implicit motives and one of six levels (Johannßen et al., 2020).

During the heated debate on the shared task, moderate critics of the task called for the organizers to 'pull the plug', others went as far as comparing the task with Eugenics in the Third Reich (even though those critics were clearly out of line with the constructive and fair debate held by most researchers).

Three main debated upon topics emerged from the debate, which will be discussed in this paper: i) IQ testing and biases, ii) forbidden research, and iii) reasons for even building such a system.

In Section 2, we will first describe the shared task in more detail and provide some background information of IQ testing in Section 3. A discourse as to why this is an ethical question and the course of the emerged Twitter shitstorm and heated debate is presented in Section 4. The three points of discussions are in Section 5 (i, IQ and bias), Section 6 (ii, reasons for building such a system), and 7 (iii, is there forbidden research). A final discussion in Section 8 concludes.

2 The role of implicit motives for the GermEval20 Task 1

Researchers have found indications that linguistic features such as function words used in a prospective student's writing perform better in predicting academic development (Pennebaker et al., 2014) than other methods such as GPA values.

The purpose of the GermEval20 Task 1 was to investigate, whether firstly, implicit motives can be classified on a human level and whether secondly, those implicit motives are sufficient for compensating flawed predictors utilized during aptitude diagnostical evaluations, such as GPA or IQ scores (Johannßen et al., 2020).

During an aptitude test, participants are asked to write freely associated texts to provided questions on shown images (such as those displayed in Figure 1). Psychologists can identify so-called implicit motives from those expressions. Implicit motives are unconscious intrinsic desires that can be measured by implicit methods (Gawronski and De Houwer, 2014; McClelland et al., 1989) Those implicit psychometrics are said to be predictors of behavior and long-term development from utterances (McClelland, 1988; Scheffer, 2004; Schultheiss, 2008).

From a small sample of an aptitude test collected at a college in Germany, the classification and regression of cognitive and motivational styles from a German text can be investigated (Johannßen and Biemann, 2019).

3 Introductionary words on IQ testing

IQ testing is a form of psychometrical testing, mostly utilized in the area of aptitude diagnostics and structural assessments. The term *intelligence* itself is debated upon, as well as those types of tests themselves (see Section 5). Nowadays, IQ testing is in an academic crisis, forming a paradigm shift towards more precise tests of isolated, single skills rather than one defining metric, dissolving misconceptions of IQ testing as being a sort of personality trait or fixed property of an individual.

Problems and questions from IQ tests vary greatly and range from the use of language, proverbs, mathematics to causalities or mechanical problems, as displayed in Figure 2.

Attention, auditory, visual and tactile perception, language, memory, and executive function all need to be considered when assessing the IQ

(Christie, 2005). One of the most comprehensive IQ tests is the Wechsler test. It samples verbal and non-verbal areas of intellectual functioning (Wechsler, 2011).

For each IQ test, the mean of all participants marks 100 testing points, with the standard deviation adding or subtracting 15 points. As a result, 68% of a population range between 85 to 115 IQ testing points.

Since IQ tests can only discriminate abilities between the percentiles from 3 percent to 97 percent of the tested population, very weak or very strong individuals can not be identified. IQ testing may only be valid by identifying and testing a truly representative peer group. The more homogeneous the peer group, the more valid the test. As humans are rich in diversity, this criterion is hardly accomplishable in practice and should thus be accounted for when interpreting scores.

Furthermore, the cultural and environmental exposures of peer groups and individuals are crucial. For many international IQ tests, only individuals, that were exposed to representative use of the English language at all stages of life may be comparable to other peers.

Aptitude diagnosticians call for IQ tests to be utilized only to identify possible child weaknesses to purposefully support them with additional educational offers. (Christie, 2005)

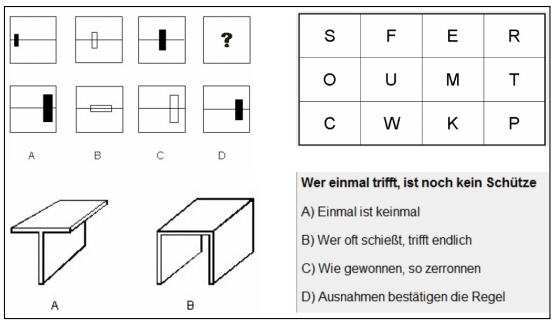

Figure 2: Different parts from an IQ test utilized at the Nordakademie. Upper left: logical comprehension, upper right: memory skills, lower left: technical comprehension, lower right: linguistic comprehension.

4 Discourse – why is this an ethical question?

When there is a need for discussing the ethical considerations of a shared task, there has to be a moral dilemma in the first place. Philosophical ethics is a branch of philosophy, which is concerned with how humans should live and what is considered to be just or unjust. Morals, on the other hand, are normative and rather connected to the activities performed by humans under the premise of ones' ethics. Whilst philosophical ethics define a set of ground truths about how to live in a just way, moralities are implications on what to act out in certain situations.

Moral dilemmas are issues of conflicts between what a person should do and should not do. In those cases, where there is neither a choice of clearly right nor wrong action in a moral situation, the dilemma becomes present (Braunack-Mayer, 2001).

As will be described in Subsection 4.1, there is a necessity for aptitude testing in nowadays human interactions, may it be that a scholar is to be chosen, a new employee or a high school grade. Psychologists face the challenge that many cognitive processes are difficult to measure and to observe. Oftentimes, psychologists have to rely on behavioral observations or questionnaires. However, any consciously given response can never reveal unconscious desires, which is why e.g .implicit motives (see Section 1) might be valuable. One downside to the use of implicit motives is that they yet have to be fully understood and researched in terms of their validity. Many metrics are yet not fully understood, even though psychologists are confident in their explanatory powers.

IQ tests have been researched for more than 100 years and are said to be among the most stable, valid, and reliable metrics in psychology (Benson, 2003). However, recent research has clearly shown that there is hardly any performed IQ testing has been done without introducing harmful biases (see Section 5). As standardized tests are the closest feasible form of aptitude diagnostics, they can hardly be completely omitted either. This is a strong moral dilemma.

This very dilemma caused for the GermEval20 Shard Task 1 so eagerly debated upon. It bears the chance of learning more about implicit motives, which even could compensate IQ testing, replacing it completely. However, any research on IQ testing-related data bears the danger of biases, pseudoscience, and misuse of approaches or results.

4.1 Aptidude diagnostics and IQ testing in the NLP research community

When it comes to working with natural language processing methods in combination with IQ tests, there are different disciplines.

One of the most broadly researched disciplines is benchmarking artificial intelligence (AI) systems by their capabilities of scoring well on human-intended IQ tests. Even though data resources for performing those AI benchmarks are limited, small, and poorly standardized (Liu et al., 2019), there are still many experiments on them. The goals of those benchmarks include advancing AI, validating AI systems, investigate intelligence testing further, and to understand better what human intelligence is.

Another discipline is more closely linked to human behavior and tries to correctly classify or human performances on IQ or comprehension tests according to the defined measures of success of the test. Different from the benchmarking, where the IQ test itself is the research object, this discipline always involves the IQ test and human performance. The shared task provides data for such studies. Another example of such a task is the Automated Student Assessment Prize (ASAP) Short Answer (SA) challenge.[4] The ASAP-SA was conducted by the Hewlett Foundation and aimed for predictions of students' grades based on short answers given by those students. Even though there were ethical discussions beforehand and during the task, the organizers specifically dismissed those and asked for methodology papers only. Regarding the ethical concerns, the organizers argued that standardized tests occupy professionals, which could, were they not manually evaluating those tests, able, to craft more sophisticated, more individual and more insightful testing procedures.

The third discipline is a meta-analysis on circumstances connected with NLP on IQ tests. Those studies include e.g. bias research, statistical evaluations, or behavioral consequences. Ethical considerations, such as this work, can be included in that discipline as well (Tsvetkov et al., 2018; Hovy and Spruit, 2016).

[4] https://www.kaggle.com/c/asap-sas/overview/description

4.2 Heated Community Discussion

On December 4, 2020, the first call for participation was released on multiple channels, one being the corpora list[5]. The original call was entitled 'GermEval 2020 Task 1 on the Prediction of Intellectual Ability and Personality Traits from Text: 1st Call for Participation'. It described the task in very technical terms with a focus on the NLP methodology.

A first reaction on the Corpora List considered that "[a]s a community, we should think carefully about whether it is appropriate to work with IQ test results as data, and what the applications of this research might be." and "In the United States, there is considerable evidence that IQ tests are racially biased"[6].

Besides the first topic of discussion, biases in IQ testing, the direct response to this introduced the second topic of discussion: "This task seems irresponsible/poorly conceived to me. Before designing such a task, I think it is imperative to consider its use cases: When and why would we want to predict IQ scores or high school grades from the text?"[7].

The discussion continued in an argumentative manner and a respectful tone. There were roughly equal amounts of supporters and opponents of the task and even though many assumptions were made – e.g. of whether the data was provided voluntarily by the aptitude test participants – quickly, panelists came to a conclusion, that too little of the underlying circumstances of the task was described in the first call, to sensibly continue the discussion.

In the meantime, the discourse continued on another channel: Twitter. As shown in Figure 3, concerns were raised by an initial tweet.

Other than the well-formulated and balanced concerns on the Corpora List, tweets are usually much shorter. Many researchers asked for more details, others drifted into speculations on not-provided details of the task's circumstances. As the tone got increasingly hostile and demands for 'pulling the plug' and starting petitions against

[5] https://mailman.uib.no/public/corpora/2019-December/

[6] https://mailman.uib.no/public/corpora/2019-December/030882.html By Jacob Eisenstein

[7] https://mailman.uib.no/public/corpora/2019-December/030883.html by Emily M. Bender

[8] A message on the short messaging service Twitter is called a tweet and can be up to 240 characters in length.

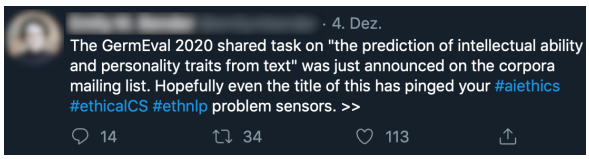

Figure 3: A first tweet carried the discussion from the NLP specific corpora list to the more visible and international social plattform Twitter, raising concerns about the shared task.[8]

it arose, the task organizers posted a request for some time to formulate a call revision. This revision was released on the 5th of December, one day after the 1st call of participation, the organizers released a statement, clarifying motivations for this task and explaining some of the task's circumstances[9]. Meanwhile, the shitstorm on Twitter had escalated up to the lowest point of comparisons to Nazi Germany and Eugenics, as displayed in Figure 4.

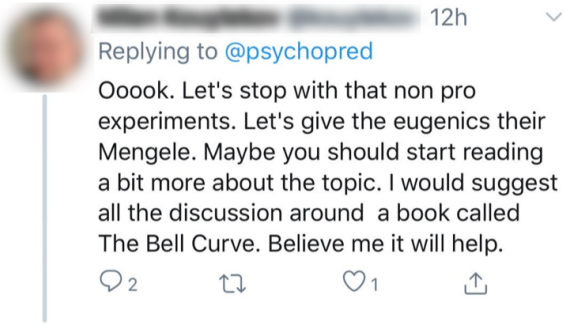

Figure 4: The so-called *shitstorm* on Twitter went so far as there were calls for petitions against the task, demands for 'pulling the plug', as well as Nazi Germany and Eugenics comparisons. However, many participants strived for a sensible, respectful and constructive discussion.

Two days after the published 1st call for participation, the third topic of discussion emerged from the discourse: forbidden research. One community researcher wrote a Medium article on the topic with the title 'Is there research that shouldn't be done? Is there research that shouldn't be encouraged?' on the 7th of December[10]. The organizers published an explanatory public reaction to the heated discussion on their accompanying task website, on a Codalab competition website, as well as on Twitter[11].

The heated discussion quickly died down after the 6th of December. The task organizers revised their task website to include more detailed information on the tasks' circumstances, ethical consideration and changed the name of the task from 'Prediction of Intellectual Ability and Personality from Text' to 'Classification and Regression of Cognitive and Motivational Style from Text', as this is by far more precise in terms of the tasks' goals (as described in Section 1).

5 IQ testing and biases

This section will explore and discuss the first of the three topics of discussions that emerged from the discourse: i) IQ testing and biases. Effects like measurable biases and a training effect will be addressed, as well as measures taken against those biases and a general discussion on systems theory.

5.1 There is a socioeconomical bias in IQ testing

Minorities can be discriminated by a biased due to unequal environmental circumstances and measurements in non-representative groups (Rushton and Jensen, 2005).

Firstly, the term *intelligence* in intelligence testing is highly misleading, as there is no well-defined concept of intelligence. Rather than *intelligence*, IQ tests are said to measure the skill of the specific tasks employed in an IQ test.

In what IQ scores measure, they are thought to be one of the best measures in the scientific field of psychology. Those skills are often, what aptitude diagnostics define as relevant for many modern skill-oriented jobs (validity). They furthermore stay relatively stable across different stages of life, starting with the age of 8, and with time (stability). The stay consistent when performed with the same setups but different observers and observees (reliability). Lastly, parts of IQ tests correlate with each other, suggesting an inference of both, environmental and genetic factors (Plomin and Deary, 2015).

[9] https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/germeval-2020-cognitive-motive/germvval2020-task1-public-statement.txt

[10] https://medium.com/@emilymenonbender/is-there-research-that-shouldnt-be-done-is-there-research-that-shouldn-t-be-encouraged-b1bf7d321bb6 by Emily M. Bender.

[11] https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/germeval-2020-cognitive-motive/germvval2020-task1-public-statement.txt

However, there are strong controversial signs of this genetic pool thought. The Flynn effect, which states that the IQ scores among the population grow by 3 points each decade, by far too fast for it to be connected to evolution, shows this. This effect is rather caused by skills acquired from environmental changes (Flynn, 1987). E.g. When investigating the development of refugees, their early environmental circumstances are suboptimal. However, later positive environmental factors can compensate for those early difficulties (Dweck, 2017).

Not so much race, but socioeconomics is thought to be the difference (Turkheimer et al., 2003). It shows that IQ potential is determined by genes, but whether this potential develops is dependent on the environment (who is rich can achieve anything). Thus, only if it was ensured that both individuals enjoyed the same good environment, an IQ score would truly say anything.

IQ tests are good measures of innate skill ability if all other factors are held steadily, which is, in fact, impossible. The differences in IQ scores across minorities and the majority, as it is present, points to a very serious issue: Inequality of opportunity. It is this socioeconomic bias, which leads to unequal opportunities especially in countries where there is a rich diversity amongst the population.

5.2 Standardized tests are trainable

As stated before, terms like *intelligence* or *cognitive ability* are misleading, when describing aptitude diagnostics and IQ testing. Those tests measure a pre-defined set of skills.

In the area of psychology, there are certain principals for experiments and procedures, that have developed over many years. One of those principals is that participants always need to know what is being tested beforehand. It needs to be clear, what results of a procedure or test are, and how those results relate to the participants.

Another principle is that most experiments and tests in psychology are only truly objective, if participants did not yet complete this particular test and do not know the exact type of testing that the whole experiment or parts of it will set up.

As soon as there is knowledge or accustoming of the underlying testing procedure, participants will most likely perform more in the direction of what they think would maximize a reward, then they would if they did not take a test yet. In other words, IQ tests are trainable. The more often participants perform IQ tests that are thought to be the norm of a given point in time (in reference to the Flynn effect), the higher participants are thought to score on those tests.

However, this implication does not just hold for the IQ tests themselves but goes further. As stated before, IQ tests do not test *intelligence* but skill. As those skills are connected with many skills required by modern jobs and with many skills trained and taught in schools – e.g. to think in abstract terms, pattern recognition, and basic math skills like the Fibonacci series), well educated and often home-schooled individuals tend to train skills more and thus perform better on *intelligence* testing procedures, even though high skills on any non-related task can not be guaranteed.

5.3 IQ Tests Discriminate Minorities

It is this bias, which leads to unequal opportunities especially in countries where there is a rich diversity among the population. Intelligence testing has had a dark history. Eugenics during the great wars e.g. in the US by sterilizing citizens [12] or in Germany during the Third Reich and Eugenics (Reddy, 2007) are some of the most gruesome parts of history.

But even in modern days, the IQ is misused. Recently, IQ scores have been used in the US to determine which death row inmate shall be executed and which might be spared. Since IQ scores show a too large variance, the Supreme Court has ruled against this definite threshold of 70 (Cooke et al., 2015). However, Sanger (2015) has researched an even more present practice of 'racial adjustment', adjusting the IQ of minorities upwards to take countermeasures on the racial bias in IQ testing, resulting in death row inmates, which originally were below the 70 points threshold, to be executed.

There is an ethical necessity to carefully view, understand and research the way intelligence testing is conducted and how those scores are – if at all – correlated with what we understand as 'intelligence', as they might be mere cognitive and motivational styles. Further valuable research can be conducted to investigate connections between other personalities. Racial biases are measurable,

[12] https://supreme.justia.com/cases/federal/us/274/200/

variances are great and many critics state that IQ scores reflect upon skill or cognitive and motivational style rather than real intelligence as it is broadly understood.

5.4 Wrong Wording: 'Cognitive ability'

As stated before, the term *intelligence* is highly misleading. Furthermore, for the first subtask, we were trying to reproduce, what is being utilized as a selection mechanism, where the IQ testing is only part of the test. High school grades are another, matched to implicit motive tests. For the second subtask, implicit motive texts were to be classified directly.

Both of those tasks aimed for researching an underlying pattern or truth to implicit motives, which have been researched and show no inherent bias. Furthermore, high school grades in Germany consist of 60% participation in class. Thus the whole term *cognitive ability* in the 1st call for participation of GermEval20 Task 1 was nonsensical and plain wrong but was revised quickly after first concerns were raised.

5.5 Relation to German socioeconomic structure

There is a broad understanding that intelligence testing is especially prone and biased towards the environments and circumstances in which they were developed. As a result, the tests designed in Western, white societies are problematic when utilized for testing richly diversified cultures (Vahidi, 2015).

Unfortunately, the German education system is known to have a strong socioeconomic bias, which leads to a vast under-representation of people with a migration background in higher education (Diehl and Fick, 2016; Fernandez-Kelly, 2015).

This, paradoxically, leads to the data of the GermEval20 Task 1 being less prone to the influence of such biases, with respect to the ground population of the underlying data. Even though there is no information on names, nationalities, or other demographic data, as it is forbidden in Germany to record personal information according to EU laws. However, e.g. pictures of graduation years indicate that there is little diversity amongst those graduates, as displayed in Figure 5.

5.6 Challenging IQ testing biases in Germany

During the long history of research on the field of IQ testing, many mistakes were made and investigated. Aptitude diagnosticians have spent decades to challenge and correct the strong socioeconomic biases, that were present in most of the earliest IQ tests. Nowadays, there are many different variants and approaches of IQ testing.

Figure 5: The exemplary 2014 year of graduation from the NORDAKADEMIE illustrates the cultural homogenity, as the vast majority of graudates are white (https://www.shz.de/lokales/elmshorner-nachrichten/lasst-die-huete-fliegen-id19354606.html). In Germany, a strongly biased socioeconomical filter is already present at the high school level.

In Germany, there is little diversity amongst private college applicants. Even though researchers at the NORDAKADEMIE try to actively challenge those socioeconomic biases by employing implicit motives, that are known to be less biased than other metrics in the field of aptitude diagnostics, the employed IQ test also accounts for the little diversity of the participant population.

The NORDAKADEMIE utilized the IST 200 R intelligence structural test by Liepmann et al. (2007), which was normalized on high school graduates[13]. Since only about a third of students attend high school in Germany, the basic population of this IQ test accounts for the little diversity of most applicants at the NORDAKADEMIE, which already experienced a socioeconomic filter. Even though this filter is a discrimination already, the employed IQ test objectively accounts for the type of the basic population that takes the test and thus challenges this bias.

[13] In Germany, the secondary school tier consists of three types of schools: The Hauptschule (practice-oriented vocational education), the Realschule (theory-oriented vocational education) and the Gymnasium (high school, preparations for pursuing a college education). Only about 30% of graduates go to college in Germany (Fernandez-Kelly, 2015)

5.7 Systems theory by Luhmann: communication is the relation of systems and none can be transmitted without noise

The systems theory by Luhmann is a philosophical and sociological communication theory, that describes agents of an environment not as instances but in their relations to other agents. Communication, according to Luhmann, is the constructing principle of an environment and not just a mere tool. An agent is understood as an autonomous part of this environment, which offers its inner structure as a matter of communication to other agents (Görke and Scholl, 2006).

However, as the channel model of communication by Shannon (1948) describes, there is no communication between agents (sender and receiver) without being obscured and disturbed by noise.

One environment or system is science. Every scientific discipline can be described as an agent in this environment. Whenever there is incomplete knowledge of the inner state of an agent, any type of communication between those systems gets obscured by noise and thus assumptions of those inner states can range from approximations to mere guesses. In any case, the assumptions are flawed.

Applied to the GermEval20 Task 1 and the ethical dilemma of IQ testing at hand, it can be stated, that since the scientific field of applied NLP does not comprehend the inner state of the scientific field of psychology and aptitude diagnostics, assumptions of the implications, limits, and effects of IQ testing from any non-psychological researcher must be viewed with caution – especially, if no correspondence has been undergone, as truth is the interaction between correspondence, consensus and consistency (Sahakian and Sahakian, 1993).

6 Resons for building a system that GermEval20 Task1 proposes

This section deals with the topic ii) of discussion: why should such a potentially dangerous system, as proposed by the GermEval20 Task 1 organizers, be build in the first place? Difficulties and possible misuses are the subjects of this section, as well as some background on the implications of researching implicit motives.

6.1 Short text classification is difficult and vague

Short text classification is a very difficult task. The most widely used method is the keyphrase extraction (Zhang et al., 2018). However, the implicit motive theory asks for annotators to examine the narratives of texts rather than single keywords or keyphrases (Scheffer and Kuhl, 2013). Thus, the most promising method for short text classification is not applicable for implicit motives texts, and therefore, it is doubtful that the mere focus on an automatic classification procedure creates valid results.

6.2 The task reflects an established practice in Europe

It can be debated upon, whether researchers should focus on theoretical tasks or if a very practical focus is legit. The NORDAKADEMIE is a university for applied sciences and the whole context of the GermEval20 Task 1 aims for researching implicit motives in the very application-oriented field of aptitude diagnostics[14].

Even if there are very good and strong arguments against aptitude diagnostics, assessment centers, the consideration of socioeconomically biased high school grades or personal job interviews, it is a very common practice in Germany and Europe to examine all of those approaches for decisions on whom to employ.

Mainly companies in Europe employ IQ tests for selecting capable applicants. In the United Kingdom, roughly 69 percent of all companies utilize IQ. In Germany, the estimate is 13 percent (Nachtwei and Schermuly, 2009).

Academia has the responsibility to research the benefits of society. Even though the organizers of the GermEval20 Task 1 do not focus on IQ testing but the implications of implicit motives, since IQ testing is part of the conducted practice in Germany and Europe, there is an academic responsibility to research its implications. Furthermore, science nowadays is called upon making efforts towards findings that are closely related to everyday society, as Bormann (2013) points out.

6.3 Purpose of the task

Especially the early reactions of the shitstorm described in Section 4.2 were shaped by misconceptions, incomplete information and misleading as-

[14] https://idw-online.de/de/news492748

sumptions what the GermEval20 Task 1 was about and what was it not about:[15]

"[...] I would worry about any research project whose organisers chose to include "prediction of intellectual ability" in the very title. Presumably a careful choice for a big research project. [...]"

This becomes apparent as some Tweets raised concerns mostly based on the headline of the 1st call rather than its content, whereas the 1st call paired with provided websites did not include much of background information either.

However, as the organizers stated in their task companion paper (Johannßen et al., 2020) and more prominently on their revised companion website: "Any research performed on this aptitude test or the annually conducted assessment center (AC) at the NORDAKADEMIE is under the premise of researching methods of supporting human resource decision-makers, but never to create fully automated, stand-alone filters"[16].

The defined goal of the GermEval20 Task 1 Subtask 1 was to reproduce a ranking of students based on the sum of z-standardized high school grades and IQ scores solemnly based on provided implicit motive texts (Johannßen et al., 2020). Those ranks were not calculated to indicate the superiority of single individuals over others. From an aptitude diagnostical view, this would not make sense. E.g. a student that might achieve a high IQ score and high German high school grades but worse math and English grades might have a higher overall *rank* compared to a student whose metrics are all above average but without any especially high ones. Yet companies might prefer someone who is above average in every aspect over anyone, that might have high grades in one subject but very low ones in other (Hell et al., 2007).

Moreover, the critics of this shared task and the organizers themselves have criticized the broad consideration of IQ scores and high school grades in Germany and the EU, as they discriminate against minorities with their socioeconomic bias.

[15] https://mailman.uib.no/public/corpora/2019-December/030896.html by Mike Scott.

[16] https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/germeval-2020-cognitive-motive.html

7 Forbidden research

This section explores the topic of discussion iii): Is there research that should not be done? Is there forbidden research? As these are questions for broad fundamental debates, we will only focus on those aspects, that appear most connected to the GermEval20 Task 1.

7.1 Knowledge is not harmless

The first general principle to acknowledge is that knowledge is not harmless. There are many examples of theoretical research being utilized for destructive follow-up research or dangerous utensils directly. Exemplarily, Alfred Nobel did not intently dynamite to be used for war, but rather for mining. Historians assume that Nobel included a peace dedicated Nobel Price in his last will is due to his invention being misused for war purposes[17].

This is an example of a so-called dual use of inventions. When inventions intentioned for civil uses is misused without the consensus of the inventor for military purposes, this is called dual use. Williams-Jones et al. (2014) describe dual use more generally as being used for good and bad either intentionally or unintentionally by the inventors.

Furthermore, the authors describe the dilemma of this dual use, as there is rarely any impactful research that could not be considered dual use. Most meaningful findings could be utilized for the good and the bad. Moreover, at times it is not even possible to imagine the negative or bad dual use of one's inventions, as further research has not been conducted yet and novel products have yet not been seen (Williams-Jones et al., 2014).

One infamous example of dual use that was not necessary imaginable is nuclear energy and its characteristics, which has to lead to a lot of scientific progress (e.g. research on cancer treatments), civilian use (e.g. nuclear energy), but also great destructions and threats (e.g. nuclear weapons) (Tucker, 2012, p. 74 ff.).

7.2 NLP can easily be misused for pseudoscience

IQ scores are prone to pseudoscientific settings and are not easily distinguishable from serious and sophisticated settings, thus masking the overall utility of IQ testing.

[17] https://www.nobelprize.org/alfred-nobel/alfred-nobels-thoughts-about-war-and-peace/

Some participants of the heated discussion of the GermEval20 Task 1 criticized this task for being "dangerously pseudoscientific". To understand, what this criticism refers to, one must first understand pseudoscience. Hansson, a Swedish philosopher, first differentiate science from pseudoscience in that scientists enjoy common *raison d'être* to provide the reader with the most epistemically warranted statements (Hansson, 2013, p. 62 ff.) by employing known and broadly respected methods for finding those statements.

Furthermore, Hansson describes the correspondence between different scientific fields and disciplines that are interconnected. No given statement violates statements made by other disciplines and fields.

As for pseudoscience, authors are mostly divided as to which characteristics define pseudoscience. However, two major characteristics appear to be agreed upon by most authors: i) Non-science posing as science and ii) doctrinal components (Hansson, 2017).

For pseudoscience to be posing as science paramount effort is undertaken to mask statements as being made with those scientific principles, even if they are not. As science offers advantages of describing true phenomena and reality, pseudoscientists strive for acceptance by readers with statements, that normally would not hold the thorough process of scientific work.

For pseudoscience to be of deviant doctrine, the pseudoscientists put sustained effort to promote standpoints different from those that have scientific legitimacy. Thus, pseudoscientists disregard major principles of scientific work, like correspondence, consensus, and consistency, as well as transparent methodology, replicability or intersubjectivity (Sahakian and Sahakian, 1993).

As for the GermEval20 Task 1, critics saw either non-scientific work being presented as scientific one or a doctrine, disregarding established methods from corresponding scientific fields, which are NLP and psychology. The main arguments for calling the shared task pseudoscience is most likely the view, that since IQ testing is viewed by many researchers as biased and unprecise, even asking for machine learning systems would be pseudoscientific. They view the methodology as not being reconcilable with established ones.

Furthermore, discussion participants criticized that a shared task holds a scientific premise, which they did not view appropriate for a task, that is – in their view – methodologically flawed.

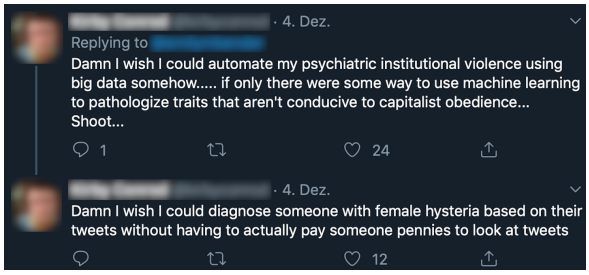

Figure 6: Critics used sarcasm to express their view of the shared task being methodologically flawed. This, paired with the scientific framing that GermEval offers lead to accusations of pseudoscience. The point of the task, to research implicit motives, appears to have been missed by some.

However, as shown in Section 6 on the point of discussion ii), many critics mistakenly assumed that the task is about building an automated system for ranking students or classifying IQ scores, whilst, in turn, it is only about researching the implicit motive theory, as the response on the corpora mailing list shows:[18]

"[...] lending legitimacy to the use of similar tools that are used as a pseudoscientific mantle to disguise (essentially) the automation of racial/ethnic/cultural discrimination and biases [...]"

7.3 Marketplace of ideas

Even in case of the criticism on the GermEval20 Task 1 setup, automation or IQ testing being legit and point to issues, there are still strong ethical arguments for educational institutions against giving in, when broadly and publicly being exposed to social media sanctions and calls for "pulling the task"[19].

One of those arguments is the *marketplace of ideas*, which was first discussed by John Stuart Mill in his 1859 book *On Liberty* (Mill, 2011). The marketplace of ideas is an analogy to the free market and assumes, that when ideas, statements, and thoughts are presented with almost perfect information – that is, on a transparent, replicable

[18] https://mailman.uib.no/public/corpora/2019-December/030885.html by Yannick Versley.

[19] https://medium.com/@emilymenonbender/is-there-research-that-shouldnt-be-done-is-there-research-that-shouldn-t-be-encouraged-b1bf7d321bb6 by Emily M. Bender.

and reliable basis –, only the truth will emerge from this available marketplace. Gordon (1997) describes the marketplace of ideas as a metaphor, where people speak and exchange ideas freely. Freely means, that there is as little interference from the government and the society as possible.

Reflected upon the GermEval Task 1, there is a violation of this philosophical and the liberal principle of freely spoken ideas. Whilst the government did not interfere with ideas presented, the society did in the way of strong social media pressure and sanctions, not discussing the idea itself but rather demanding the idea to be stopped without professional discourse (at least on Twitter, as discussions on the corpora mailing list, were mostly argumentative[20]).

At times, ideas, scientific interests, and projects might provoke criticism due to a Zeitgeist, even if they are legit and worth exploring. Darwin (1859) was heavily criticized and even got his novel work on his theory of evolution banned. Criticism on his theory not only arose from religious organizations but even from respected and well established fellow researchers.[21] However, if the marketplace of ideas would have been applied to Darwin, his theory and findings would have been openly discussed with all forms of scientific research, arguments, and findings, leaving it to the audience and research community to determine, whether his theory holds for the moment. If his ideas, however, would have been banned, as suggested by some critics of the GermEval20 Task 1, there would not have been an open debate. Furthermore, if Darwin's theory was utterly wrong, it would not have been able to compete and thus vanished.

7.4 Knowledge cannot be restrained

As Grashon (1983) describes, multiple researchers announced to leave science, after having discovered the knowledge of isolating DNA fragments for the first time. They feared that this discovery would lead to political and social pressure. One of those scientists even formed a group, categorically pressuring any scientific work on this genetic field. Nonetheless, DNA sequencing has continued to be researched.

There are implications, that – at least basic – research discoveries can not be fully prevented or stopped, as the so-called *multiple discovery* or *simultaneous invention* principle calls for them to be made. This multiple discovery principle is the hypothesis, that most discoveries are made independently by multiple scientists at the same time, often internationally. The Nobel price committee often recognizes this hypothesis by rewarding multiple scientists who, at that time, did not collaborate directly.

This hypothesis is thought to be observable, since discoveries, theories and scientific tools enable practicing scientists of a field to now make discoveries. As the circumstances are ideal in an internationally spread research community, simultaneous inventions are made possible. One example is radar technology, which was discovered by multiple countries independently and at the same time (Galati, 2015). Thus, many believe the suppression of scientific progress is not possible.

On the other side, Martin (1978) argues, that a development of science and technology emerging from that science independent from the thoughts and desires of single scientists and pressure from society are historically incorrect. In his article, the author argues with selected examples, namely nuclear power, food additives, transport policy, genetic engineering, and automation – all of which are characterized as technologies, having emerged from basic research and having experienced pressure and concerns from the research community and society. What the author does not argue about, is the value of basic research itself. He states that the path of scientific and technological development is not usually predictable beforehand. Furthermore, Martin notes that concerns over scientific and technological development has almost always to do with *applications and implications* for the wider society.

At times, the research could have assumed what negative impact a discovery or invention could have on society, as Nobel, which invented the dynamite mainly for supporting mining, could have imagined the use for military purposes. Nonetheless, the individuals utilizing dynamite to build weaponry are rather to blame than Nobel himself, even if he greatly regretted, that his discovery was used for such[22].

[20] https://mailman.uib.no/public/corpora/2019-December/

[21] https://en.wikipedia.org/wiki/Reactions_to_On_the_Origin_of_Species

[22] https://www.nobelprize.org/alfred-nobel/alfred-nobels-thoughts-about-war-and-peace/

7.5 Pushing scientists out of academia

Whilst the US has spent 4,545.7 Mio. dollars (Pece, 2020) in research and development (R&D) of computer sciences and mathematics, the US Department of Defense possessed a R&D budget of 52,973.3 Mio. dollars, which is more than 40% of the total US R&D budget. Some of the most influential advancements in computer science has been researched *behind closed doors* for military purposes such as the RSA cryptosystem, which was already invented by the GCHQ four years before the later patented peer-reviewed method[23] or the predecessor of the internet, the ARPANET, which was developed by the U.S. Airforce in 1969 (O'Neill, 1995).

Some private companies possess comparably large R&D budgets as well: Alphabet, the parent company of the Google corp. spent 26,018 Mio. dollars on R&D [24]. Even though the most recent scientific advancements were made open-sourced and have been peer-reviewed, such as the bidirectional encoder representations from transformers (BERT, (Devlin et al., 2019)) and Tensorflow 1.0.0 (Fujita et al., 2017, p. 564), earlier developments, such as the Google PageRank algorithm, which was kept hardly reproducable, despite even patents describing the basic procedure (Lindberg, 2008).

One causality and risk of violations of the marketplace of ideas is that researchers, which experience pressure, might leave the public academia to pursue research in the private sector, which does not necessarily publish research to be reviewed, discussed, and criticized by the public. This could lead to knowledge monopolies, as well as fraudulent or misconducted research.

This is further reflected by the recent development, that influential technology companies have caused a so-called AI brain drain, meaning, that many countries experience the emigration of AI researchers. A national brain drain is observable from the public research sector and academia to private firms due to higher salaries, greater funding, and at times more academic freedom (Kunze, 2019).

8 Discussion

Whenever basic research leads to new technologies and applications, there is a risk of misuse. As humans nowadays produce a vast amount of digitally available textual resources, research of NLP applications could quickly lead to questionable and possibly dangerous results. The GermEval20 Task 1 has rightfully sparked a heated debate upon the ethical considerations of this task, as it not only involves NLP methods but furthermore aptitude diagnostics, psychometrics, and IQ testing – all of which can and have been misused.

However, as we have shown in this paper, the three main topics of discussions, i) IQ testing and biases, ii) reasons for building such a system and iii) forbidden research, have positively be evaluated in terms of their ethical indications.

IQ tests are very prone to biases. As the data was collected from a small university of applied sciences in Germany, the peer groups are heterogenous, no score can be reverse-engineered from the available data and as the main point of the task is not to automate IQ testing but research implicit motives, we believe the discussion to have lost track of what the task is truly about.

This also leads to the second topic of discussion. We have shown that in Germany, high schools already function as destructive socioeconomical filters that discriminate against minorities. Implicit motives have shown to be by far less biased and more neutral. If they were better understood and validated, aptitude diagnostics could finally move away from bias-prone metrics as high school grades or IQ scores.

Lastly, we have shown calls to forbid most research topics to not only be misleading, but harmful. In a marketplace of ideas, only truth can emerge. The past has additionally shown that progress is hardly containable. Moreover, public shaming and condemnation of research ideas could lead to them moving to the private sector, which already has a lot of innovative power without the necessity to present, discuss and have ideas criticized by peers – all of which are some keystones of scientific work.

We view this ethical discussion as partly unobjective but have also seen a valuable discourse from most of the participants. It is right to view any research project critically. However, it is always important to closely investigating what a research idea is truly about and to honor scientific freedom, as forbidding certain ideas puts this freedom at stake.

[23] https://www.wired.com/1999/04/crypto/
[24] https://abc.xyz/investor/

9 Acknowledgements

Firstly, we want to thank the NLP community, which participated in the important discourse of this task. Especially the fellow researchers from the corpora list gave valuable input and points of view.

Furthermore, we want to thank Emily M. Bender, which first raised reasonable concerns of the task and extended the discourse to Twitter, as well as to Medium, where she formulated two posts, summarizing and evaluated main reasons for those concerns. Also, we would like to thank her for participating in the KONVENS 2020 Ethics and NLP panel. Additionally, we want to thank Twitter users, which also took place in the discourse, providing even more and broader insights in social implications of the task.

We want to thank Michele Loi, who wrote a profound ethical assessment of the task provided us with objective and neutral ethical arguments and agreed to participate in a constructive NLP + society session alongside Emily M. Bender and Dirk Johannßen on the SWISSTEXT & KONVENS conference.

We want to thank the SWISSTEXT & KONVENS committee for supporting our task, neutrally investigating the ethical impacts of it and provided expertise and aid in determining the task's risks and chances.

We want to thank our colleagues and lab members, that supported us with advice and help in keeping an overview and objective point of view.

Last but not least, we want to thank the participants of the shared task for staying open-minded and interested in the research, providing founded empirical evidence of whether such a task is solvable and discussing its impacts on the research community and society.

References

Etienne Benson. 2003. Intelligent intelligence testing. *Monitor of Psychology*, 34(2):48–49.

Lutz Bornmann. 2013. What is social impact of research and how can it be assessed? A literature survey. *Journal of the American Society for Information Science and Technology*, 64:217–233.

Annette Joy Braunack-Mayer. 2001. What makes a problem an ethical problem? An empirical perspective on the nature of ethical problems in general practice. *Journal of Medical Ethics*, 27(2):98–103.

Deborah Christie. 2005. Introduction to IQ testing. *Psychiatry*, 4:22–25.

Brian K. Cooke, Dominque Delalot, and Tonia L. Werner. 2015. Hall v. Florida: Capital Punishment, IQ, and Persons With Intellectual Disabilities. *Journal of the American Academy of Psychiatry and the Law Online*, 43(2):230–234.

Charles Darwin. 1859. *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life*, volume 1859. London: John Murray.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, MN, USA. Association for Computational Linguistics.

Claudia Diehl and Patrick Fick. 2016. Ethnische Diskriminierung im deutschen Bildungssystem. In Claudia Diehl, Christian Hunkler, and Cornelia Kristen, editors, *Ethnische Ungleichheiten im Bildungsverlauf: Mechanismen, Befunde, Debatten*, pages 243–286. Springer Fachmedien, Wiesbaden.

Carol S. Dweck. 2017. From needs to goals and representations: Foundations for a unified theory of motivation, personality, and development. *Psychological Review*, 124(6):689–719.

Patricia Fernandez-Kelly. 2015. The Unequal Structure of the German Education System: Structural Reasons for Educational Failures of Turkish Youth in Germany. *Spaces & flows : an international journal of urban and extraurban studies*, 2:93–112.

James Flynn. 1987. Massive IQ gains in 14 Nations: What IQ tests really measure. *Psychological Bulletin*, 101:171–191.

Hamido Fujita, Ali Selamat, and Sigeru Omatu. 2017. *New Trends in Intelligent Software Methodologies, Tools and Techniques: Proceedings of the 16th International Conference Somet 2017*. Ios Press, Washington, DC, USA.

Gaspare Galati. 2015. A Simultaneous Invention – The Former Developments. In *100 Years of Radar*, 1st ed. 2016 edition, pages 55 – 77. Springer, New York, NY, USA.

Bertram Gawronski and Jan De Houwer. 2014. Implicit measures in social and personality psychology. *Handbook of research methods in social and personality psychology*, 2:283–310.

Elliot S. Gershon. 1983. Should science be stopped? The case of recombinant DNA research. *The Public interest*, 71:3–16.

Jill Gordon. 1997. John Stuart Mill and the "Marketplace of Ideas". *Social Theory and Practice*, 23(2):235–249.

Alexander Görke and Armin Scholl. 2006. Niklas Luhmann'S theory of social systems and journalism research. *Journalism Studies - JOURNAL STUD*, 7:644–655.

Sven O. Hansson. 2013. *Defining pseudoscience and science*. University of Chicago Press, Chicago, IL, USA.

Sven O. Hansson. 2017. Science and Pseudo-Science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, summer 2017 edition. Metaphysics Research Lab, Stanford University.

Benedikt Hell, Sabrina Trapmann, and Heinz Schuler. 2007. Eine Metaanalyse der Validität von fachspezifischen Studierfähigkeitstests im deutschsprachigen Raum. *Empirische Pädagogik*, 21(3):251–270.

Dirk Hovy and Shannon Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Dirk Johannßen and Chris Biemann. 2019. Neural classification with attention assessment of the implicit-association test OMT and prediction of subsequent academic success. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS*, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Dirk Johannßen, Chris Biemann, Steffen Remus, Timo Baumann, and David Scheffer. 2020. GermEval 2020 Task 1 on the Classification and Regression of Cognitive and Motivational style from Text. In *Proceedings of the GermEval 2020 Task 1 Workshop in conjunction with the 5th SwissText & 16th KONVENS Joint Conference 2020*, pages 1–10, Zurich, Switzerland (online). German Society for Computational Linguistics & Language Technology.

Rolf-Torsten Kramer, Werner Helsper, Sven Thiersch, and Carolin Ziems. 2009. *Selektion und Schulkarriere: Kindliche Orientierungsrahmen beim Übergang in die Sekundarstufe I*. Studien zur Schul- und Bildungsforschung. VS Verlag für Sozialwissenschaften.

Lars Kunze. 2019. Can We Stop the Academic AI Brain Drain? *KI - Künstliche Intelligenz*, 33(1):1–3.

Detlev Liepmann, André Beauducel, Burkhard Brocke, and Rudolf Amthauer. 2007. *Intelligenz-Struktur-Test 2000 R*. Hogrefe Verlag, Göttingen, Germany.

Van Lindberg. 2008. *Intellectual Property and Open Source: A Practical Guide to Protecting Code*. O'Reilly & Associates, Sebastopol, CA, USA.

Yusen Liu, Fangyuan He, Haodi Zhang, Guozheng Rao, Zhiyong Feng, and Yi Zhou. 2019. How Well Do Machines Perform on IQ tests: a Comparison Study on a Large-Scale Dataset. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6110–6116. International Joint Conferences on Artificial Intelligence Organization.

Han Maas, Kees-Jan Kan, and Denny Borsboom. 2014. Intelligence Is What the Intelligence Test Measures. Seriously. *Journal of Intelligence*, 2:12–15.

Brian Martin. 1978. Can scientific development be stopped? *Australian Science Teachers Journal*, 24(1):65–70.

David C. McClelland. 1988. *Human Motivation*. Cambridge University Press.

David C. McClelland, Richard Koestner, and Joel Weinberger. 1989. How do self-attributed and implicit motives differ? *Psychological Review*, 96(4):690–702.

John Stuart Mill. 2011. *On Liberty*. Cambridge Library Collection - Philosophy. Cambridge University Press.

Jens Nachtwei and Carsten C. Schermuly. 2009. Acht Mythen über Eignungstests. *Harvard Business Manager*, (04/2009):6–10.

Judy O'Neill. 1995. The role of ARPA in the development of the ARPANET, 1961-1972. *Annals of the History of Computing, IEEE*, 17:76 – 81.

Deniz S. Ones, Stephan Dilchert, Chockalingam Viswesvaran, and Jesús F. Salgado. 2017. Cognitive ability: Measurement and validity for employee selection. In *Handbook of Employee Selection, Second Edition*, pages 251–276. Taylor and Francis.

Christopher Pece. 2020. Federal R&D Obligations Increase 8.8% in FY 2018; Preliminary FY 2019 R&D Obligations Increase 9.3% Over FY 2018. Technical Report 20-308, National Science Foundation.

James W. Pennebaker, Cindy K. Chung, Joey Frazee, Gary M. Lavergne, and David I. Beaver. 2014. When Small Words Foretell Academic Success: The Case of College Admissions Essays. *PLOS ONE*, 9(12):e115844.

Robert Plomin and Ian J. Deary. 2015. Genetics and intelligence differences: five special findings. *Molecular Psychiatry*, 20(1):98–108.

Ajitha Reddy. 2007. The eugenic origins of IQ testing: Implications for post-Atkins litigation. *DePaul L. Rev.*, 57:667.

Detlef H. Rost. 2009. *Intelligenz: Fakten und Mythen*. Beltz, Weinheim, Germany.

John P. Rushton and Arthur R. Jensen. 2005. Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law*, 11(2):235–294.

William S. Sahakian and Mabel L. Sahakian. 1993. *Ideas of the Great Philosophers*. Barnes & Noble, New York, NY, USA.

Robert M. Sanger. 2015. IQ, Intelligence Tests, 'Ethnic Adjustments' and Atkins. SSRN Scholarly Paper ID 2706800, Social Science Research Network, Rochester, NY, USA.

David Scheffer. 2004. *Implizite Motive: Entwicklung, Struktur und Messung [Implicit Motives: Development, Structure and Measurement]*. Hogrefe Verlag, Göttingen, Germany.

David Scheffer and Julius Kuhl. 2013. *Auswertungsmanual für den Operanten Multi-Motiv-Test OMT*. sonderpunkt Verlag, Münster, Germany.

Frank L. Schmidt and John E. Hunter. 1998. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124(2):262–274.

Oliver C. Schultheiss. 2008. Implicit motives. *Handbook of personality: Theory and research*, pages 603–633.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technology Journal*, 27(3):379–423.

Yulia Tsvetkov, Vinodkumar Prabhakaran, and Rob Voigt. 2018. Socially Responsible NLP. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 24–26, New Orleans, LA, USA. Association for Computational Linguistics.

Jonathan B. Tucker. 2012. *Innovation dual use and security: Managing the risks of emerging biological and chemical technologies*. MIT Press.

Eric Turkheimer, Andreana Haley, Mary Waldron, Brian D'Onofrio, and Irving Gottesman. 2003. Socioeconomic Status Modifies Heritability of IQ in Young Children. *Psychological science*, 14:623–8.

Siamak Vahidi. 2015. Intelligence Testing and Cultural Diversity: Pitfalls and Promises | The National Research Center on the Gifted and Talented (1990-2013). Library Catalog: nrcgt.uconn.edu.

David Wechsler. 2011. *WASI-II: Wechsler abbreviated scale of intelligence*. NCS Pearson, San Antonio, TX, USA.

Bryn Williams-Jones, Catherine Olivier, and Elise Smith. 2014. Governing 'Dual-Use' Research in Canada: A Policy Review. *Science and Public Policy*, 41:76–93.

Yingyi Zhang, Jing Li, Yan Song, and Chengzhi Zhang. 2018. Encoding Conversation Context for Neural Keyphrase Extraction from Microblog Posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1676–1686, New Orleans, LA, USA. Association for Computational Linguistics.