Can Conceptual Engineering Reduce Implicit Bias?

Changing Language (to Change Thought) to Change Behavior

Bernt Ivar Barkved

MAPSYK330: Master's Thesis in Social & Cognitive Psychology
The Faculty of Psychology
University of Bergen
May 2023

Supervised by Henrik Berg[1]

[1] Department of Clinical Psychology & Center for the Studies of the Sciences and Humanities, University of Bergen.

Abstract

A scoping review was conducted to investigate the potential conceptual engineering to reduce implicit bias. The review investigates whether using conceptual engineering to reduce gendered language can reduce implicit bias. The empirical evidence is inconclusive. While there is a clear correlation between gendered language and implicit bias, language change interventions do not seem to have the effects that conceptual engineering advocates claim. A systematic review on the effects of conceptual engineering in reducing implicit bias is called for. Furthermore, the evidence of using conceptual engineering to reduce implicit bias is promising enough to be included in diversity programs, like implicit bias training.

Keywords: Conceptual Engineering, Gendered Language, Implicit Bias

Word count: 12057

Sammendrag

Det ble gjennomført en kartleggingsoversikt for å klargjøre potensialet til å bruke conceptual engineering for å redusere implisitt bias. Omfanget ble innsnevret til å kartlegge om å redusere kjønnet språk kan revideres til å redusere implisitt bias. Bevisene er blandet og ufullstendige. Det virker å være en klar sammenheng mellom kjønnet språk og implisitt bias, men det er uklart om språkforandinger har de ønskede effektene som filosofer har argumentert for. Det anbeles å gjøre en systematisk kartlegging av hvilke effekter språkforandringer kan ha på å redusere implisitt bias. Det anbefales også å bruke conceptual engineering som en del av mangfoldsprogrammer som implisitt bias trening.

Nøkkelord: Conceptual Engineering, implisitt bias, kjønnet språk

Antall ord: 12057

Acknowledgements

I would like to thank Karin Lillevold, Annika Rødeseike, Audun Syltevik, and Jon Arild Aakre for helpful advice and comments. I also want to thank my friends and family for their support. Finally, I want to thank my supervisor, Henrik Berg, for invaluable comments and support.

Table of Contents

1. Introduction

Jessica Nordell (2021) begins her book, *The End of Bias*, with the example of Ben Barres, a neurobiologist who underwent gender reassignment as an adult. In 1995, without knowing what *transgender* was until a year after, he had a double mastectomy, and subsequently began hormone treatment. Barres was unsure how the scientific community would react. Taken by surprise, he experienced being treated with much more respect by his colleagues when perceived as a man, than as a woman. Until the gender reassignment, he had not realized the amount of disrespect and sexism he had previously been exposed to. He had not been taken seriously in discussions; all his ideas, contributions and authority were devalued. One scientist, unaware of Barres being transgender, even commented: "Ben gave a great seminar today_but then his work is so much better than his sister's," (Nordell, 2021, p. 2). (When, of course, his sister was himself!) "It is not that Barres never encountered barriers and bias, he told me of his career before his transition," Nordell (2021, p. 3) writes, and quotes Barres, "«It was just that I didn't see it.»"

Presumably, most of the colleagues did not act discriminatory against Barres on purpose. Unintentionally, they were guilty of *implicit bias*, which is the topic of this thesis. Implicit bias is, in brief, to "act on the basis of prejudice and stereotypes without intending to do so," (Brownstein, 2019). It is acknowledged that implicit bias causes considerable harm in many facets of society: employment discrimination (e.g. Krieger & Fiske, 2006); disparities in medical treatment (e.g. Chapman, et. al., 2013); unjust criminal justice (e.g. Richardson & Goff, 2013); criminal behavior, e.g. the graphic police murder of George Floyd (Goff, 2021), and so on. However, it might be even more widespread than most people know. Neither the victim nor the assailant need to be aware of discriminatory behavior (as illustrated in the above example). Most strategies to reduce implicit bias target the individual to change themselves. Some examples are simple awareness of one's biased behavior (Gonzalez, et. al., 2021) and engaging in mindfulness (Kang, et. al., 2014; Magee, 2016). However, it is difficult to overcome implicit bias by only focusing on individual awareness and intervention. Prejudice scholars have called for a shift of focus from individuals to the sociocultural world to get to the root of the problem (Adams, et. al., 2008). There is a surge in *Implicit bias training*, where companies, governments and universities mandate people to attend courses to be trained and educated on implicit bias (Pankey, et. al., 2018), as means to reduce implicit bias. The problem is that there is little evidence that implicit

bias training has an effect (Applebaum, 2019; FitzGerald, et. al., 2019; Hagiwara, et al., 2020; Jackson, 2018; Kim & Roberson, 2022; Noon, 2018).

There is an additional option to reduce implicit bias, heeding the shift to the sociocultural world, that ought to be explored. This option is to *engineer language*. It entails intentionally changing the meaning of words and concepts in order to achieve certain goals. This trend has been named *Conceptual Engineering* (CE) and has become a sub-topic in philosophy and beyond. CE is a method that can be applied on anything, from everyday language (e.g. Appiah, 2018; 2022; Haslanger, 2000; 2012) to philosophy (e.g. Clark & Chalmers, 1998), economics (e.g., Herfeld, 2022), geography (e.g. Casati, 2022) and psychology (e.g. Churchland, 2021; Sunstein, 2021; Tanesini, 2022; Tremain, 2021). The method reflects the theory of linguistic relativity that the words people use affect the world and how they think about it. Research in linguistics supports such a corresponsive relationship between language, thought and behavior (Everett, 2013; Reines & Prinz, 2009). One aspiration in CE is that changing concepts and words can help improve the sociocultural world (Cappelen, 2018). I refer to *language change* throughout as: a change or "aim at a change of language, e.g., by introducing [words and] concepts that were engineered for a certain purpose," (Löhr, 2022, p. 836).

The effect that CE can have on the influence of implicit bias has not been explicitly addressed in the prejudice literature. It is this gap in the literature I aim to explore by answering the following research question:

> *Research Question:* What does the prejudice literature convey about CE's effectiveness in reducing implicit bias?

There is a normative question of moral and legal permissibility of engineering language. I assume that it is permissible to engineering language (cf. Lohr, 2022) for purposes like reducing implicit bias. There are two other aspects to the research question. One aspect is the effectiveness of CE. If CE does not lead to change in language, it cannot be an effective strategy for reducing implicit bias. This challenge has been discussed extensively (e.g., Nimtz, 2021; Jorem, 2021; Queloz & Bieber, 2021), and the literature indicates that language change is a common outcome (cf. Nimtz, 2021; Simion & Kelp, 2020). The other aspect is whether the proposed changes in language will have the desired effects (cf. Fischer, 2020). Can language change (qua CE) reduce implicit bias?

The research question needs specification. CE contains a multitude of approaches to language changes (e.g. Nimtz, 2021; Pinder, 2021; Simion & Kelp, 2020; Haslanger, 2012). Therefore, a hypothesis was created:

Hypothesis: Engineering language to reduce gendered language reduces implicit bias.

Gendered language was chosen over, for example, racialized language because of brevity and a greater extent of literature on the topic. Psycholinguists distinguish between three categories of gender use in language: grammatical gender, natural gender and genderless language (Corbett, 1991; Gygax, et. al., 2019, p. 3). In *grammatical gender* languages, like French, every noun is either masculine or feminine. In *natural gender* languages, like English, individuals are referred to by the gender pronouns he/him and she/her. In *genderless languages,* like Finnish and Turkish, neither nouns nor pronouns have gender. By *gendered language(s)* is meant grammatical gender or natural gender or combinations of the two.

The hypothesis concerns language's effect on implicit thought and behavior either (indirectly) through awareness or (directly) implicit thought. Engineering language can cause *awareness* of how one ought to speak to lessen implicit bias and discriminatory behavior. If one is informed not to say a word because it is transphobic, for instance, this might suffice to use a different vocabulary. Engineering language can affect implicit bias *implicitly*. A change of language, and differences in languages themselves, can affect people implicitly to be more tolerant. For example, there is some evidence pertaining to foreign language acquisition increasing intercultural tolerance (Gojkov-Rajić & Prtljaga, 2013). The research question is, then, dependent on behavioral change by means of language via explicit and implicit thought. Therefore, two premises need to be established to answer the research question. First, the premise that implicit or unconscious thought guides behavior, which is the essence of implicit bias. Second, the premise that language influences implicit thought. If language guides thought, and thought guides behavior, language can guide behavior through thought. Thus, *linguistic relativity* (i.e., the view that language influences the way we think) is directly tied to the research question.

Philosophers argue that engineering language to *de-gender* it can reduce stereotypes about genders and race, also implicitly (Dembroff & Woda, 2018; Ritchie, 2021). This claim is empirical. To address the research question and hypothesis, I carried out a scoping review on the

effects of de-gendering language on implicit bias. *Scoping reviews* are "preliminary assessment of potential size and scope of available research literature," (Grant & Booth, 2009). As this is a scoping review, with the purpose of locating gaps and connections, one hypothesis should suffice to give credence to the research question. I do not assume to exhaust the literature on the topic of CE and implicit bias, and there are other potential avenues in which CE can aid in the ousting of implicit bias. For example, engineering language to manipulate perceived norms to make people behave according to the norm regardless of what they think about the norm (Cialdini, et. al., 2006; Nimtz, 2021). There might also be ways of engineering language that facilitate implicit bias. For example, the attempts by some of the members of the Russian government to connect LGBTQ ideals with nazism. Neither of these matters can be adequately covered here.

Academics are putting together a unifying effort to use language as a tool for thought and as a tool to *change thought* (Isaac, 2023). The culmination of this ongoing effort is CE. The aim of this thesis is to gauge the potential of CE as a framework for reducing implicit bias. The thesis is a preliminary report on the role that language change could have on general strategies for reducing implicit bias. If the research question is supported, a practical question emerges about whether language change should be a part of implicit bias training, and other types of diversity training. Reviewing the literature on whether implicit bias can be reduced by engineering *gendered* language can give preliminary answers to that question and satisfy the aim of this thesis.

1.1 Implicit Bias

Compare Humans, or Homo Sapiens, with the imaginary Homo Economicus (Econs, for short) (see, e.g., Kahneman, 2011). Econs are perfectly rational. Humans are not. For a long time, researchers in behavioral economics assumed humans to be rational. The *expected utility theory*, modeling the decision making of perfectly rational agents, has arguably been the dominant theory since its conception in 1944 (von Neumann & Morgenstern, 1944/2007). Richard Thaler (2016, p. 9), amongst others, has emphasized the need to acknowledge the existence and relevance of irrationality in human-beings:

> We don't have to stop inventing abstract models that describe the behavior of imaginary Econs. We do, however, have to stop assuming that those models are accurate descriptions of behavior, and stop basing policy decisions on such flawed analyses. And we have to start paying attention to those *supposedly irrelevant factors*.

The relevance of human behavior in decision making was first accentuated by Daniel Kahneman and Amos Tversky (1979), in what they called *prospect theory*. The aim of prospect theory was to describe the actual behavior of humans, including those *supposedly irrelevant factors*. Kahneman and Tversky (1974; 1979; 1982) demonstrated, through a series of controlled studies, that humans are not perfectly rational agents. Humans make decisions based on heuristics, or mental shortcuts, at the expense of accuracy. Heuristics are the results of evolution, and often very useful, for instance, to avoid irrelevant information. However, heuristics can sometimes lead to severe and systematic errors (Kahneman & Tversky, 1974, p. 1124). Implicit bias is one such error caused by heuristics (see, Griffin, et. al., 2012, for an historical overview). Another example is the *the conjunction fallacy* (Kahneman & Tversky, 1982), which is illustrated by the following case:

> Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
>
> What is more probable?
> 1. Linda is a bank teller?
> 2. Linda is a bank teller and is active in the feminist movement.

The majority of participants chose Option 2. Option 2 is less likely given that the probability of two events occurring in conjunction is, by default, less likely than one of them occurring alone.

Heuristics can lead to a dissociation between the reflective and the automatic mind. The dissociation occurs between values and action. Consider an analogy. One way to reduce substance abuse is simply to pay people not to take drugs (Redish & Regier, 2015). One explanation for why payment works is that being offered money to stay off drugs puts people *out of* habitual thinking, and *into* deliberate thinking (Redish & Regier, 2015). The former is mainly controlled by the basal ganglia and cerebellum - associated with automatic responses. Deliberate thinking, contrarily, mainly uses the prefrontal cortex - associated with creativity and second-order thoughts (Redish & Regier, 2015). It seems that there are two different processes at work, one automatic and one reflective and deliberate. The automatic and reflective brain are often referred to as System 1 and System 2 thinking, respectively (Stanovich,1999; Stanovich & West, 2000; Kahneman, 2011). System 1 works automatically and quickly, with little effort and awareness. System 2 works slower, is conscious and logical, and requires more effort.

Dissociation between values and action occurs because System 1 is in control. Categorization is another process controlled by System 1.

Heuristics are mental shortcuts that humans use to, amongst other things, categorize. *Categorization* is "the process by which objects, events, people, or experiences are grouped into classes on the basis of (a) characteristics shared by members of the same class and (b) features distinguishing the members of one class from those of another," (American Psychological Association (APA), 2023b). Categorization, like heuristics, are often useful. Phylogenetically, differentiating between dangerous and harmless animals has been essential for human survival. Nevertheless, categorization often leads to discrimination (Vaughan, et. al., 1981). Barring a very few exceptions like drug reactions, it is not useful to distinguish humans by race (Appiah, 2018). In fact, the concept of RACE arguably was *invented* in the 17th century to justify enslavement of Africans (Appiah, 2018). (Concepts are designated in SMALL CAPS.) The invention of RACE has created groups based on an arbitrariness like variation in skin pigmentation.

In addition to implicit bias, there are three modes of category representation that are relevant: *prejudice,* psychological essentialism (*essentializing*, for short) and *stereotype*. Prejudice is "a negative attitude toward another person or group formed in advance of any experience with that person or group," (APA; 2023f). Essentializing is the tendency to assume underlying essence that unifies members of that group (Neufeld, 2022). Essentializing and prejudice are therefore dependent on each other. Stereotype is "a set of cognitive generalizations (e.g., beliefs, expectations) about the qualities and characteristics of the members of a group or social category," (APA, 2023d). The phenomena of *in-group bias and out-group homogeneity* illustrates all three category representations combined.

In-group bias is the tendency to favor one's own group. Michael Billig and Henri Tajfel (1973) carried out a study where they assigned participants groups based on preference of a painting. A control group was not assigned into groups. The participants were then asked to give real money to other participants. Billig and Tajfel (1973) found that participants gave more money to the group they were assigned. No difference was found in the control group. In-group bias can, in other words, be created by arbitrarily placing people in groups based on aesthetic preference. Ethnicity and sex are much more ingrained, and the in-group bias therefore stronger. Out-group homogeneity is humans' tendency to see their own group (White Christians, for example) as diverse, but an out-group (Muslims) as homogenous. A Muslim who has committed

a hate crime is rarely, if ever, described as a disturbed individual. A disturbed individual is, on the other hand, exactly the description given to Anders Behring Breivik, the terrorist of the 2011 Utøya Massacre (Juergensmeyer, 2020). A study by Andreas Olsson and Elisabeth Phelps (2004) showed White Americans and Black Americans photos of one Black face and one White face. They conditioned their study subjects to fear both faces with a mild electric shock to induce Pavlovian conditioning. It succeeded. When the participants saw faces associated with shock, more sweat was measured. They then measured whether their own ethnicity would ease the distinguishing of the response. The result showed that White Americans showed faster distinguishing of fear to a White face than a Black face, and Black Americans showed the opposite effect. They postulated that this might explain why a terrorist act by an in-group may be forgotten faster than if it is by a member of an out-group. Essentializing, prejudice and stereotyping are the mechanisms behind in-group bias and out-group homogeneity. It is apparent that it sometimes occurs without conscious awareness. System 1 has taken control, and the processes involved are implicit.

Implicit attitude is "a relatively enduring and general evaluative response of which a person has little or no conscious awareness," (APA), 2023a). Implicit attitudes might result in bias. *Implicit bias* can be defined as "to sincerely assert one thing, while at the same time acting contrary to the espoused attitude," (Schwitzgebel, 2010). One example is to declare that all ethnicities and genders are equal, while preferring the opinions of your own ethnic group and gender. Implicit bias was coined in 1995 by Mahzarin Banaji and Anthony Greenwald, in an article on implicit social cognition. They found that social behavior was not just guided by conscious control, which was widely assumed at the time. Behavior was also guided by implicit cognition of past experience, like prejudice and stereotype, resulting in implicit bias, or unintended discrimination:

> […] the signature of implicit cognition is that traces of past experience affect some performance, even though the influential earlier experience is not remembered in the usual sense—that is, it is unavailable to self-report or introspection. (Greenwald & Banaji, 1995, p. 4).

In a review of implicit bias, more than 20 years later, Greenwald and Calvin Lai (2020) accentuated implicit bias' substantial impact on business, education, law, medicine and political science. In short, automatic preferences guide the decisions of humans (Banaji & Greenwald, 2013, p. 54).

There are several strategies for reducing implicit bias. The research question of this thesis, regarding what the literature conveys about the effectiveness of CE in reducing implicit bias, is important to unearth regardless of the effectiveness of other strategies to reduce implicit bias. Different strategies can be complementary. The effects of implicit bias training are, however, varied and unknown (Applebaum, 2019; FitzGerald, et. al., 2019; Hagiwara, et al., 2020; Jackson, 2018; Kim & Roberson, 2022; Noon, 2018). Many companies use, for instance, some form of *diversity training* for their employees (Nordell, 2021, p. 93). One facet of diversity training is implicit bias training. Diversity training is rarely tested systematically. Thus, the actual effects are not yet known. The research that has been done so far shows mixed results. One meta-analysis, covering hundreds of companies with diversity training initiatives, concluded that attitudes towards other groups changed slightly, for a while, but later returned to normal (Bezrukova, 2016). Another meta-analysis found that mandatory diversity training could cause undesired effects over time (Dobbin & Kalev, 2016). For example, the probability of Asian American men and women becoming managers was reduced by almost 5 percent, and for Black Women 9 percent. The study addressed many initiatives, and not only diversity training. In addition, implicit bias training is just one facet to diversity training.

In addition to being difficult to reduce, at least in the long term, implicit bias is difficult to study. There are two main reasons for this. First, implicit bias is difficult to measure. Implicit bias is usually measured with snapshot studies which do not measure the phenomenon over time. Nor can snapshot studies capture the interactive nature of bias. Jason Okonofua (et. al., 2016, p. 393) argues that "[implicit bias] arises not solely from either teachers or students, but from both acting together and perceiving and misperceiving one another." Second, some researchers argue that implicit bias needs more precise conceptualization. Bertram Gawronski (et. al., 2022), for instance, argues that the phenomenon of implicit bias is not equivalent to the bias being measured. This is because, similar to the first point, measurements cannot capture the phenomenon precisely. The most common measurement of implicit bias is the Implicit Association Test (IAT).

1.1.1 The Implicit Association Test

The IAT has produced a large body of evidence of implicit bias (Hummert, et. al., 2002). "The Implicit Association Test (IAT) provides a measure of strengths of automatic associations.

This measure is computed from performance speeds at two classification tasks in which association strength influences performance" (Greenwald, Nosek & Banaji, 2003, p. 197). The idea is that the more time participants spend in associating two categories, the less those categories have in common for the participants. There are IATs for sex, ethnicity, religion, disability, age, height and weight. Here we consider the *Race IAT*, since it was the first to launch.

For the Race IAT, participants are asked to sort White and Black faces with pleasant and unpleasant words. Sorting White faces with pleasant words quicker than Black faces with pleasant words is a sign of *automatic white preference* and a sign of hidden race bias (Dasgupta, et. al., 2000). Independent of ethnicity, nearly 75 percent of participants in America share automatic white preference (Banaji & Greenwald, 2013, p. 47). In a meta-analysis on Race IAT they found that, for example, automatic white preference correlated moderately with selection of presidential candidates in the U.S.: John McCain or Barack Obama (Greenwald, et. al., 2009). Race IAT proved to predict discriminatory behavior better than any previous method in the study on prejudice, like self-report (Banaji & Greenwald, 2013, p. 47).

Thierry Devos and Banaji (2005) created an IAT to measure the stereotype 'Asian Americans = foreign' (for Americans). The result showed that symbols such as the dollar were associated more with White Americans than Asian Americans - even by Asian Americans themselves (associating their own group as foreign). Devos and Banaji (2005) went further, and chose well-known Asian Americans paired with well-known White Europeans. The result showed that White *Europeans* were more associated with the concept AMERICAN than Asian Americans. The stereotype that Americans are white (and male) persists beyond citizenship.

Even though women work as much as men do, the prevalent gender career associations are 'Male = Career' and 'Female = Family' (Nosek, Banaji & Greenwald, 2002). Researchers found that 75 percent of male respondents exhibited those automatic gender stereotypes, while it was *80* percent for female respondents (Nosek, Banaji & Greenwald, 2002). One does find, however, that this tendency is less likely the younger one is. Laurie Rudman and Jessica Heppen (2003) found that the more strongly women associate their romantic partner with chivalric rescuers, the less they aspire for status and power themselves. This suggests that implicit bias might also be a hindrance for people exhibiting the biases. Several studies found that self-undermining stereotypes play a significant role in conserving one's own disadvantage (Jost & Banaji, 1994; Jost, Banaji & Nosek, 2004). This accounts for women staying out of

high-paying jobs, but also men often avoiding caregiver jobs. Another study found that women who hold the strongest stereotype of 'science = male' are least likely to major in science, while men who hold the same stereotype are most likely to major in science (Smyth, Nosek & Greenwald, 2009). This also extends beyond the United States. Boys outperform girls in STEM courses to a much greater extent in countries with a stronger 'science = male' stereotype (Nosek, et. al., 2009).

The assertion that thinking affects performance is found elsewhere. Two examples are *enclothed cognition* and the *Pygmalion effect*. Enclothed cognition is when wearing clothes characteristic of a profession encourages students to adopt traits, and pursue a career in that profession (Adam & Galinsky, 2012). Wearing a white lab coat can improve STEM test scores. When the participants wore artist smocks, conversely, the scores were 50 % lower (Adam & Galinsky, 2012). The *Pygmalion effect* is high expectation leading to better performance, and vice versa (Rosenthal, 1973). In a series of experiments, Rosenthal and colleagues (1973; with Babad, 1985; with Jacobson, 1968) told teachers that certain children were "growth spurters" based on some tests. The tests were fake, and the children designated as "growth spurters" were chosen at random. Based only on this information, teachers showed higher expectations of these children, and treated them accordingly. The higher expectation, and difference in treatment, caused significant improvement in student achievements.

The IAT has received criticism for having measurement errors. Thomas Carpenter (et. al., 2022) found that the IAT has weak test-retest reliability, containing considerable noise, concluding that a single IAT is inadequate to estimate trait bias. Aggregating across multiple IAT's can, however, improve its validity. Klaus Fiedler (et. al., 2006) identifies five major problems with the IAT, two of which are the difficulty of interpreting scores and susceptibility for participants to fake their answers. Melanie Steffens (2004) found that IAT is less susceptible to faking than the Big-Five personality test, but that IAT is not immune to faking. There are, then, difficulties surrounding measuring implicit bias. However, the phenomenon is seen everywhere, and there is little doubt of its pervasiveness in society. It can, therefore, be concluded that unconscious thought *sometimes* affects the way people act. The first premise for the research question is, then, affirmed: implicit thought can guide behavior. The next premise in need of affirmation is linguistic relativity: whether language can affect implicit thought.

1.2 Linguistic Relativity

The question of whether CE can reduce implicit bias is to ask whether language can affect behavior. For language to have an effect on behavior it often has to go through thought. There are two different hypotheses stating that language influences thought, differing only in degree: *linguistic determinism* and *linguistic relativity* (Everett, 2013; Li, 2022). The former is the hypothesis that language *determines* the way people think about the world. The latter is the hypothesis that language *influences* the way people think about the world. In order to affirm the research question, that CE is effective in reducing implicit bias, only one of them has to be true. The focus here will therefore be on the more benign linguistic relativity, formerly the Sapir-Whorf hypothesis.

The view that language is nothing more than expressions of thought was the dominant position in philosophy for a long time (Locke, 1690/2008; Fodor, 1975). This has been a popular belief in psychology as well, at least in the second half of the 20th century (Pinker, 1984; Piaget & Inhelder, 1948/1967). The relationship between thought and language was considered one-directional: thought precedes language. The opposite idea, that language can precede thought, was popularized in the early 20th century by Franz Boas, Edward Sapir and Benjamin Whorf. Lack of empirical support caused its downfall, but recent empirical research has revived the view (Reines & Prinz, 2009). Caleb Everett (2013), in his book *Linguistic Relativity*, compiles an array of evidence that language affects thought, ranging from space, time, gender and color terminology. For example, Mandarin and English speakers differ on conceptions of time (Boroditsky, 2011). Mandarin speakers are more likely to think about time vertically (Boroditsky, et. al., 2011). Research on linguistic relativity has also been criticized for not being able to rule out confounding factors. English speakers and Mandarin speakers are part of very different cultures, which could be the reason for the difference in conception of time–and not the languages.

Maria F. Reines and Jesse Prinz (2009) argue, however, that linguistic relativity offers the most promising interpretation of some recent research on color, numbers and spatial relations. They offer two manifestations of linguistic relativity. The first is that “languages influence psychological processes because they instill *habits of thought* that lead us to think in certain ways” that we would not have thought of without it (Reines and Prinz, 2009, p. 1028, my emphasis). Reines and Prinz (2009) argue that such habits is the best explanation of, for example,

bilinguals continuing to show the biases of their mother tongue when speaking a different language (Boroditsky, 2011). The second indication is that "languages influence psychological processes because they lead us to organize the world into categories that differ from those we would discover without language," (Reines & Prinz, 2009, p. 1029). They argue that this manifestation can explain multiple empirical results. For example, the syntactic and semantic structure of quantification-neutral nouns in the Yucatec Mayan language is different from the English language, and the speakers of this language demonstrated significant differences in unitizing (Lucy & Gaskins, 2003).

Philip Wolff and Kevin J. Holmes (2011) did not find support for language overwriting pre-existing conceptual distinction, but did find support for language *augmenting* certain types of thinking. Wolff and Holmes (2011) conclude that, "Although the literature on linguistic relativity remains contentious, there is growing support for the view that language has a profound effect on thought." In sum, there is enough indication that language has a significant effect on thought. The extent of this effect remains contentious. Regardless of the size of the effect, however, there is a justifiable connection between thought being influenced by language. This is enough to show that it is feasible that engineering language can reduce implicit bias.

1.3 Conceptual Engineering

CE is the name given to the process of assessing and improving our concepts - or conceptual schemes - to achieve certain goals, be they political, social, theoretical or otherwise (Isaac, Koch & Nefdt, 2022: 1). This includes changing a word-meaning pair by either removing it (e.g. slurs), inventing/capturing it (e.g. SEXUAL HARASSMENT) or revising/replacing it (e.g. RAPE) (Koslow, 2022).

Two of the most popular projects associated with CE are Sally Haslanger's (2000; 2012) proposals to ameliorate gender and race concepts in order to promote social justice and equal rights, and Andy Clark & David Chalmers' (1998) proposal to revise the concept of BELIEF to make the concept more unified and useful. Haslanger's proposal to ameliorate WOMAN, specifically, is to include subordination into the definition of the concept of woman. The motivating idea is that changing the concept will give attention to the systematic subordination of women, and therefore, fight it. Haslanger (2000, p. 46) writes, "I believe it is part of the project of feminism to bring about a day when there are no more women (though, of course, we should

not aim to do away with females!).” Clark and Chalmers (1998) developed an idea of active externalism where objects within the environment function as part of the mind. They argue that beliefs are not limited to the insides of our heads, but can be extended to external things like rearranging letter tiles to prompt word recall on Scrabble; mentally rotating objects on the screen to make them fit in Tetris; using pen and paper to perform long multiplication; hypothetical neural implants. Their suggested revision is for the concept of BELIEF to include *external* beliefs.

Some projects classified as CE outside of philosophy are The International Astronomical Union’s (IAU) revision of the concept PLANET to improve the categorization of the solar system (see, e.g., Egré & O’Madagain, 2019); the improvement of the concept GENE to allow for a more context-sensitive usage in biology (see, e.g., Brigandt, 2010); Carl Linnaeus’ classification of whales from being in the extension of FISH to being in the extension of MAMMAL (see, e.g., Sainsbury, 2013); APA (2013) first introducing, then removing ASPERGER'S SYNDROME as a distinct mental disorder, including it instead as part of the Autism Spectrum. Let us consider the planet case and the Asperger’s case in more detail.

The International Astronomical Union (IAU) decided in 2006 to revise the concept of PLANET. In the 20th century there was no accepted definition of what a planet was, it only had nine canonical instances. Call this concept $\text{PLANET}_{\text{OLD}}$. In the early 21st century, however, similar celestial objects to Pluto were discovered near Neptune. This created a conundrum for IAU, whereupon they outlined two different candidates to replace $\text{PLANET}_{\text{OLD}}$:

> $\text{PLANET}_{\text{DRAFT}}$: a celestial object X that (a) orbits the sun, (b) is sufficiently large for its own gravity to have formed it into a sphere
>
> $\text{PLANET}_{\text{NEW}}$: a celestial object X that (a) orbits the sun, (b) is sufficiently large for its own gravity to have formed it into a sphere, and *(c) has cleared its neighbourhood of debris,* (Pinder, 2020: 3).

The two proposals only differ with regard to criterion (c). Selecting $\text{PLANET}_{\text{DRAFT}}$, Pluto would still be a planet, but so would Eris. Following a vote, the IAU selected $\text{PLANET}_{\text{NEW}}$.

In 2013, Asperger’s syndrome was removed from the Diagnostic and Statistical Manual (DSM) because of inconsistent application of the disorder, and similarities between individuals with Asperger’s and individuals with autism. The American Psychiatric Association (2013) decided that the clinical term should be removed from the DSM, and replaced with Autism Spectrum Disorder. In 1994, the concept of ASPERGER’S SYNDROME was introduced with the purpose of prompting researchers to identify potentially different subgroups of autism (Klin &

Volkmar, 2003). Both the introduction and the removal of ASPERGER'S SYNDROME are examples of CE.

There are three aspects that make these projects into CE: normativity, utility and intention. First, one has to take a normative approach to traditional philosophical questions, not asking what our concepts *actually mean*, or have meant in the past, but what our concepts *should* mean (normativity). It does not matter what PLANET meant in the past, it matters what it should mean. Secondly, while traditionally philosophers have attempted to accurately describe concepts' usages, conceptual engineers want to actively *use* concepts for particular purposes (utility). Concepts are tools to employ, assess and improve upon. To a certain degree, concepts are means to an end. Notice that the IAU *voted* on the best proposal. It was not that one proposal *tracked nature* better than the other, it was whatever could serve the IAU's purposes best. Thirdly, concepts, and the relationships between concepts, change all the time. This is sometimes referred to as *conceptual change*, other times as *semantic drift* (Cappelen, 2018, p. 30). Understanding this process - what makes concepts change and why - is difficult. The relevant question here is not *how* or *why* these changes occur, but if conceptual changes can be *effected*; to what degree can we intervene, and guide and influence conceptual changes?

It is often assumed that CE is employed in the attempt to *improve X*. *X* can be anything from better understanding the world, making the world a better place, improving language, solving problems, etc. More specifically, *X* can be to normalize marriage between people of the same sex by allowing same-sex people to marry, calling it "marriage" and not "gay marriage". *X* can also be to lessen stigmatization caused by the association with a virus by changing the names of the variants of the coronavirus from geographical places to letters from the Greek alphabet. However, CE does not *entail* improvement: A suggestion to change a concept can cause a negative change. Nor does CE *assume* good intentions. The Russian Government's attempts to change the concept of NAZISM, so as to include the notion of LGBTQ, is also an example of CE.

There are, however, some problems with CE. The two biggest problems have to do with implementation and feasibility. The two problems overlap, so I find it helpful to view the first as *practical* problems and the second as *theoretical* problems with CE. The first, examples of practical problems, have received a lot of attention in CE, and has to do with that few, if any, proposals for changing concepts are actually getting implemented. One key problem is to reach out to enough (and the right) people to actually make the suggestion for the changes to the

concepts happen. If the proposals to engineer concepts won't be adopted, there is a question of what the purpose of the method is to begin with? The question of feasibility, which has received less focus, but is more interesting for our purposes, has to do with factors that have to do with the human psyche: can we actually convince people to change their language (and what does that entail)? In a similar vein as the implementation problem, if convincing people to change their language verges on impossibility, what is the value of the method (and philosophy, more generally)?

Not much is written about the question of feasibility in CE, connecting it to psychology. Eugene Fischer (2020), Edouard Machery (2021) and Allison Koslow (2022), all drawing on empirical research, are the exceptions. Fischer (2020:12) questions regular people's conceptual control, referring to *the salience bias*: people's tendency to focus on information that grabs our attention, which is a function of *exposure frequency*. People will tend to keep the meaning of a word that they are used to instead of the new meaning. Machery (2021:2) states that concepts have a particular psychological nature as *attractors*: the mind is drawn to think with these concepts, and efforts to replace them are therefore unlikely to succeed. Koslow (2022:15) asserts that people tend to avoid loaded words, have a difficulty of understanding proposed changes to words and that meanings have a tendency to persist even when words to use them perish.

Relevant for the hypothesis supporting the research question of this thesis, Haslanger's suggestion to ameliorate the concept of WOMAN is *not* a proposal to de-gender language. In fact, Haslanger's proposal has received so much attention, that the main focus in CE has been on conceptually negotiating the concept of WOMAN. An engineering of the concept has already taken place, at least in academic feminism, where it is widely accepted that WOMAN include trans-women (Stock, 2022). For the project of de-gendering language, it is suggested that engineering the concept of marriage, to include people of all sexes, can "promote the idea that gendered husband/wife roles are inessential to marriage and this encouraging heterosexuals to adopt more egalitarian relationship models (Pollock, 2019, p. 89). Moreover, Robin Dembroff and Daniel Wodak (2018) have explicitly argued that people should stop using gendered pronouns. One of their motivations is to avoid essentialism. The details of how to de-gender language is, however, bound to be complicated. Katherine Ritchie (2021), for instance, contends that certain *nouns* can encourage essentialist thinking. The use of nominalized terms, such as "is a female", instead of "is female" or "is a Black" instead of "is Black" cause a "psychological

propensity to essentialize" and invite thoughts that there are further shared, stable, and explanatory features of the group or kind. They may also bring to mind stereotypes about the social kind (Ritchie, 2021, p. 461).

2. Method

To address the research question, a scoping review was conducted. A scoping review is a type of literature review. Chris Hart (1998, p. 13) has provided a definition:

> A literature review is "the selection of available documents on the topic, which contain information, ideas, data and evidence written from a particular standpoint to fulfill certain aims or express certain views on the nature of the topic and how it is to be investigated, and the effective evaluation of these documents in relation to the research being proposed.

A scoping review fulfills these descriptions, but the focus is on a "preliminary assessment of potential size and scope of available research literature," (Grant & Booth, 2009). This scoping review was conducted on the following research question: What does the prejudice literature convey about CE's effectiveness in reducing implicit bias? There are many aspects and theories one could have included when scrutinizing this research question, including phraseology, gender studies and more. The purpose of this thesis was not, however, to provide an exhaustive analysis or review of the topic. The purpose of this scoping review was twofold: to assess what data there is in the prejudice literature on the research question and its hypothesis, and to gauge the potential of CE's effectiveness on reducing implicit bias. The focus of this literature search was therefore on gathering the following data: empirical findings and theories from the prejudice literature in psychology of language change affecting implicit bias. The search was narrowed down to giving supporting or opposing evidence to the hypothesis of which the research question depends:

Engineering language to reduce gendered language can, in turn, reduce implicit bias.

2.1 Concerns and Limitations

To demonstrate the strengths and weaknesses of this scoping review, it is helpful to compare it to a systematic review. A "systematic review differs in that it attempts to uncover "all" of the evidence relevant to a question and to focus on research that reports data rather than concepts or

theory," (Aromataris & Pearson, 2014, p. 54). In this review, the examination of the literature was less extensive. A scoping review was chosen over a systematic review for two reasons. First, a systematic review is the gold standard. The demands for completion are strict in a systematic review, and outside the framework of this master's thesis. Second, scoping review fits the aim of this thesis, which was to to *gauge* the potential of CE as a framework for reducing implicit bias. According to Munn (et. al., 2018, p. 1):

> Researchers may conduct scoping reviews instead of systematic reviews where the purpose of the review is to identify knowledge gaps, scope a body of literature, clarify concepts or to investigate research conduct. While useful in their own right, scoping reviews may also be helpful precursors to systematic reviews.

There are two concerns with scoping reviews that I want to highlight. The first concern is that some relevant literature might have been overlooked. Of course, one might omit relevant literature in a systematic review too. Nonetheless, omittance is more likely in a scoping review since the examination is less exhaustive. I hope to have alleviated this concern with the presentation of the literature review process. The second concern is that focused literature reviews are more vulnerable to biases than systematic reviews. Because this literature review has been less extensive, it is more prone to *confirmation bias*. Confirmation bias is to look for information that is consistent with one's beliefs. This is a genuine risk that I hope to assuage by having actively looked for literature that goes against the research question. Moreover, because my hypothesis was conducted from philosophical arguments, I might have been more prone to *the framing effect* and a lack of objectivity. Framing effect is when one is affected by the way the information is presented, instead of *what* the information is about. I do not think, however, that the framing effect was prevalent as these hypotheses were tested with a literature search. Objectivity is a real concern, however. I wrote a master's thesis in philosophy on CE, where I got my idea for this project. Based on the literature, CE seemed to be in an ideal position to reduce stereotypes and biases, like implicit biases. However, my thesis in philosophy was a *criticism* of naive optimism in CE. One of the arguments was that the proposals for language change lacked an assessment of psychological factors; whether language changes would have the effects that the philosophers proposed. I come, then, from both an optimistic and a skeptical viewpoint of CE, thinking that it can have a positive impact, but also skeptical of its psychological feasibility.

2.2 The Literature Review Process

I started off with a basis in philosophy, writing a master's thesis in philosophy at the University of Bergen on the scope and feasibility of CE (Barkved, 2022). I got the idea that it could be beneficial to consider how CE could provide a useful framework for reducing implicit biases. This is when I began my literature search. The databases used for the review were google scholar and APA psychINFO.

I first set out to find empirical research on the effects of language change on implicit bias. This provided both too many results, and a lack of relevant results. I therefore created a hypothesis that could support the research question. The idea of considering gendered language comes from recent attention in media and politics, and philosophical arguments (Dembroff & Wodak, 2018; Ritchie, 2021).

I then set out to find empirical research on the hypothesis. I began by finding out whether gendered language affects perception negatively (implicitly or not). Put differently, is gendered language in need of changing? Second, I set out to find whether gender-*neutral* language affects people's perception, and if so, how. Is there any point to changing from gendered language to gender-neutral language? Third, I set out to find out if there were any comparative studies of gendered language and gender-neutral language. Fourth, I set out to find data on the difference in countries with gendered language and non-gendered language. Articles considered relevant to the research question and hypothesis was selected based on the following search words: "gendered language", "genderless language", "gender-fair language" or "gender-neutral language" in combination with "perception", "implicit bias", "unconscious bias", "unintentional bias", "unexamined bias", "bias", "stereotype", "essentialism/essentializing" or "prejudice".

3. Results

Several findings suggest that gender and ethnicity bias are present in texts like letters of recommendations (Filippou, et. al., 2019) and even Artificial Intelligence like ChatGPT (Gosh & Caliskan, 2023). There are also indications of language becoming less biased as a consequence of interventions like implicit bias training. There is less empirical research on whether changes to language have an effect on implicit bias.

3.1 Hypothesis: Reducing Gendered Language Reduces Implicit Bias

Research on gendered language mostly focuses on three categories: grammatical gender, natural gender and genderless language (Corbett, 1991; Gygax, et. al., 2019, p. 3). French is an example of a grammatical gendered language, English an example of natural gendered language and Finnish and Turkish are examples of genderless languages. Most languages do not fit perfectly with these categories, however, and researchers often add a category or two to compensate and gather more data. Pascal Gygax (et. al., 2019, p. 1), for example, adds two categories: "languages with a combination of grammatical gender and natural gender […] and genderless languages with few traces of grammatical gender and genderless languages." German and Norwegian are instances of the first, having grammatical gender, but also a third neutral pronoun. Examples of genderless languages with traces of gendered language are Orya and Basque (Gygax, et. al., 2019, p. 4), where a "few gendered forms appear in nouns with gender suffixes or gendered adjectives or verbal forms."

Research supports the claim that discriminatory language leads to implicit bias, even to explicit bias. For example, Ashford, et. al., (2018) found that "Terms such as "substance abuser" and "opioid addict" have shown to elicit greater negative explicit bias." There is also plenty of research supporting the claim that *gendered* language leads to implicit bias. Nancy Murdoch and Donelson Forsyth (1985) conducted two studies on reactions to gender-biased language. They found that,

> (1) generic phrasings were perceived to be somewhat biased and sexist, (2) designation and evaluation stereotyping was perceived to be extremely biased and sexist, and (3) neutral alternatives were judged to be appropriately nonsexist (Murdoch & Forsyth, 1985, p. 39).

Gendered language is associated with lower rates of female labor (Gay, et. al., 2013; Gay, et. al., 2018). Lewis Davis and Megan Reynolds (2018) examined the relationship between gendered language and educational gender gap, and found that individuals speaking a gendered language is strongly associated with the gender gap in educational attainment:

> Our results are consistent with the idea that gender distinctions in language increase the salience of traditional gender roles in the mind of the speaker and contribute to unequal social outcomes across genders. As is well known, the gender gap in education has been shrinking globally over the past several decades, in part due to the empowering effect of economic development (Duflo, 2012). Our analysis poses a cautionary counter-point to this trend, suggesting that some portion of educational gender inequality is linked to highly stable linguistic structures and, thus, may persist even as countries develop (David & Reynolds, 2018, p. 48).

Sara Koeser and Sabine Sczesny (2014) presented participants with a collection of arguments regarding gender-fair language and masculine generics. Results showed that participants were more likely to "change their language behavior more in the direction of gender-fairness when they had been exposed to arguments," (Koeser & Sczesny, 2014, p. 548). In another study, participants were asked to read a text in gender-fair form, with a control group reading a text in gendered form, and they found that gender-fair language increases the prominence of women in the mind (Xiao, et. al., 2022).

In a study conducted on college students' perceptions of gender, it was found that "people with negative attitudes toward transgender individuals perceive greater difficulty in using gender-inclusive language," (Patev, et. al., 2019). Dobbin and Kalev (2018) found that voluntary training is significantly more beneficial than mandatory training. 80 percent of corporations with diversity training make it mandatory (Dobbin & Kalev, 2018, p. 52). Two studies found that if a diversity program is introduced with external motives (i.e. avoiding lawsuit), participants are more resistant to change. To the contrary, when aligned with internal motivations, like management needs, there was a change (Legault, et. al., 2011). In other experiments, White people showed resentment when exposed to external pressure to control prejudice against Black people (Legault, et. al., 2011). Participants responded with an *increased* bias unless evaluating the intervention as voluntary (Legault, et. al., 2011). One study affirms that diversity training can be counterproductive and increase stereotypes by making them more cognitively accessible (Macrae, et. al., 1994).

In countries where the dominant language is more gendered is found to correlate with lower rates of female participation in labor and credit markets (Gay, et. al., 2013; Mavisakalyan, 2015). Countries with gendered language have shown to be more likely to have gender quotas (SantacreuVasut et al., 2013). Steven Samuel (et. al., 2019) conducted a systematic review of 43 studies on gender and linguistic relativity investigating whether grammatical gender assignment "rubs off" on concepts themselves. For example, they questioned whether Italian speakers would conceptualize beds as more masculine than speakers of other languages because bed is masculine in Italian. Samuel (et. al., 2019, p. 1767) concluded that,

> […] support was strongly task- and context-dependent, and rested heavily on outcomes that have clear and equally viable alternative explanations. We also argue that it remains unclear whether grammatical gender is in fact a useful tool for investigating relativity.

Andrea Bender (et. al., 2016) found that the gender congruency effect on allegorically used nouns was driven by the association of nouns with personifications rather than by their grammatical gender. Another study found that bilingual speakers exhibited intraspeaker relativity in semantic representations, indicating that gender does not have a conceptual, nonlinguistic effect (Kousta, et. al., 2008).

There are two interventions of note. Stanford University created the *Elimination of Harmful Language Initiative*, a website consisting of a list of offensive and harmful words that people were encouraged to follow. The results were counterproductive, and the initiative was concluded. This was the statement by the chief information officer (Gallagher, 2022):

> The primary motivation of this initiative was always to promote a more inclusive and welcoming environment where individuals from all backgrounds feel they belong. The feedback that this work was broadly viewed as counter to inclusivity means we missed the intended mark.

Lotta Rajalin began an experiment to provide equal opportunities for girls and boys in a preschool in Sweden, first gathering video tapes of the pupils' behavior (Nordell, 2021, p. 262). She found was that children were not the biggest problem, the teacher's where. Teachers treated boys and girls differently. For instance, teacher's comforted girls more than boys when they cried, and told girls, but not the boys, to be quiet when they were being impulsive (Nordell, 2021, p. 263). Over the next several years, the teachers altered their behavior. They let boys cry, and girls be impulsive; they flipped the genders of the characters in stories; encouraged children to play together regardless of gender; they referred to the children by their name, instead of by pronoun; replacing the gendered pronouns with a neutral pronoun "hen", common in Nordic countries; talking to children about gender issues, and more (Erdol, 2019). The experiment was met with much hostility. Rajalin received hate mail, the building was graffitied and public critics "condemned a dystopian school that was brainwashing the children and eliminating the concepts of "male" and "female"," (Nordell, 2021, p. 263). The result showed that those concepts were not eliminated. Tuba Erdol (2019) and Kristin Shutts (et. al., 2017) compared the gender-neutral school with children in other schools, and they both concluded mixed results, but that the intervention had an overall positive effect on gender stereotypes:

> […] children attending the gender-neutral preschool scored lower on a gender stereotyping measure than children attending typical preschools. Children at the gender-neutral school, however, were not less likely to automatically encode others' gender. The findings suggest that gender-neutral pedagogy has moderate

> effects on how children think and feel about people of different genders but might not affect children's tendency to spontaneously notice gender, (Shutts, et. al., 2017, p. 1).

A final study measured the public reaction in Sweden when the neutral pronoun, "hen", was introduced (Gustafsson Sendén, et. al., 2015). Reviewing the process from 2012 to 2015, they found that Swedish people were originally negative to the word in 2012, but already in 2014 there was a significant shift to more positive attitudes, and the use of the word increased. The researchers attributed time as the main factor, and concluded:

> […] new words challenging the binary gender system evoke hostile and negative reactions, but also that attitudes can normalize rather quickly. We see this finding very positive and hope it could motivate language amendments and initiatives for gender-fair language, although the first responses may be negative.

Their results demonstrated that people might change their opinions about something they were immediately aversive towards.

4. Discussion

4.1 Hypothesis: Reducing Gendered Language Reduces Implicit Bias

The hypothesis is: *Engineering language to reduce gendered language can, in turn, reduce implicit bias*. The first question that is relevant is whether gendered language needs to be reduced; is there a connection between gendered language and implicit bias? This is the topic of most of the prejudice literature. Demonstrating that discriminatory language influences implicit bias is one argument for why removing or revising these words might reduce implicit bias. For example the finding that substance abuse and opioid addicts elicit greater negative explicit bias (Ashford, et. al., 2018). In order words, implicitly biased language seems to cause the implicit bias to become explicit. Murdoch and Forsyth (1985) concluded that people have strong reactions to gender-biased language. The problem is not necessarily when gender-biased language is overt, and the participants are aware of it, but rather when people are *not* aware. Several studies have found a correlation between gendered language and injustice or unfairness. For example, lower rates of female labor (Gay, et. al., 2013; Gay, et. al., 2018). The problem with these studies is that it is difficult to control for confounding factors. In addition, there are studies that conclude that gendered languages are not *necessarily* the reasons for gender congruency: bilingual speakers exhibited intraspeaker relativity in semantic representations, indicating that

gender does not have a conceptual, nonlinguistic effect (Kousta, et. al., 200); gender congruency effect on allegorically used nouns was driven by the association of nouns with personifications rather than by their grammatical gender (Bender, et. al., 2016). In fact, a systematic review of 43 studies on gender and linguistic relativity, of which the research question of this thesis is strongly connected, concluded that there are "equally viable alternative explanations" and that it is "unclear whether grammatical gender is in fact a useful tool for investigating relativity," (Samuel, et. al., 2019, p. 1767). Another way of putting this latter claim is that it is unclear whether grammatical gender is a useful tool for investigating the way that language affects thought, which is what the connection between gendered language and implicit bias depends on. However, the link between grammatical gender and linguistic relativity might be worth investigating, not for figuring out the way language affects thought, but to get practical results; to reduce bias. Following linguistic relativity, if people speak a more gendered language, this might lead to greater gender distinctions in the mind of the speaker (Davis & Reynolds, 2018, p. 46). When gender distinctions are highlighted in the mind, this might lead to more pronounced gender roles in society (Boroditsky, et. al., 2003). So far, researcher's cannot conclusively prove that there is a strong corresponsive relationship between gendered language and implicit bias, but the manifold of IAT tests indicated that there is. In sum, a correlation is found between gendered language and implicit bias, if not a causal relation - which is difficult to prove in any study (Davidson, 1967).

The next question of relevance to investigate the hypothesis is whether there are differences in countries with gendered language contra countries with or gender-neutral language. There are strong indications for an affirmative answer to that question. There are lower rates of female labor participation (Gay, et. al., 2013; Mavisakalyan, 2015) and a greater likelihood of gender quotas (SantacreuVasut et al., 2013) in countries with gendered language, and these are just two studies of many to find such correlations. These studies suffer from the same problems as the ones above, however, in that the countries have major cultural differences that the researchers cannot exclude as causing the gender-differences. In addition, the systematic review by Samuel (et. al., 2019) is discerning for the claim that gendered language is what causes differences in implicit bias, as seen by an unfair difference in jobs and gender quotas. The same conclusion is therefore made for the second question of the hypothesis as well: a correlation is found between countries with gendered language and implicit bias.

The third question that is relevant to investigate the hypothesis is whether interventions including language change reduces implicit bias. There are few comparison studies of gendered language versus non-gendered language as a consequence of intervention. The best evidence of such interventions working are the two studies from Sweden. The preschool teacher's that changed their own behavior, including flipping the genders of the characters in stories, talking to children about gender issues and stop using pronouns by either replacing the gendered pronouns with a neutral pronoun "hen" or referring to children by their name (Erdol, 2019). These measures had remarkable effects, while at the same time not eradicating the concepts of male and female, as critics feared. Results showed that the concepts of gender were alive and well, but that the children were less obsessed by gender, and it altered and removed stereotypes of how people of different genders are *supposed* to be. In other words, it eliminated *essentialism*, to believe that certain properties were essential to the different genders. "They saw boys and girls, but they stereotyped them less," (Nordell, 2021, p. 264). The study was conducted by teachers, however, with little scientific grasp of intervention techniques and confounding factors. The teachers were the conductors of the study, and participants in it. However, the data was gathered over several years, with data from 16 different documents, and researchers measured the effects, contrasting them with other children (Erdol, 2019; Shutts, et. al., 2017). Moreover, there were several interventions, and not solely to reduce gendered language, and it is difficult to know how much of the effects one should attribute to de-gendering language.

The other intervention study in Sweden was exactly about de-gendering language, and its effects. In 2012, Sweden introduced a third gender-neutral pronoun, "hen", as an addition to the already existing pronouns for she "hon" and he "han". Gustafsson Sendén (et. al., 2015) measured the effects from the implementation in 2012 until 2015. What they found was a strong aversive reaction when hen was first made official, but given time more it shifted to more positive attitudes and increase of the use of the word. The societal effects of this change remains unclear, however. Sweden has a very liberal and modern culture, and is considered a fairly equal society between men and women. It is also a place where one suspects these changes come more easily about. Norway is another country that has adopted the use of "hen", and made this implementation official in 2022. It remains to be seen whether the addition of "hen" will have the desired effects in these countries, however, and lead to a more inclusive and less implicit bias.

The last question that is relevant to support the hypothesis is whether people would accept the language changes. The reactions from initiatives like the *Elimination of Harmful Language Initiative*, the introduction of the gender-neutral pronoun "hen" in Sweden and the experiment to provide equal opportunities for girls and boys in a preschool in Sweden all show strong aversive reactions. The Stanford University's *Elimination of Harmful Language Initiative* (Gallagher, 2022) is one attempt to use CE to reduce gendered language to reduce, amongst other things, implicit bias. This Initiative was met with backlash and criticism, and was soon concluded. Since the initiative was canceled, the possible effects are not known. It seems to have failed for much of the same reasons that the recent attempts to alter the language in Roald Dahl's books: it is a change in status quo and people feel forced to do something they do not want to do. Research shows that voluntary training is much more successful than mandatory training (Dobbin & Kalev, 2018; (Legault, et. al., 2011)), and this initiative probably felt mandated to many. In addition, the initiative went too far too fast. They created a list of, what felt to many like, *banned* words and the reasons for the banning was not entirely obvious. Researchers might understand that changing words like 'blind review' to 'anonymous review' might be beneficial in reducing implicit bias because it unintentionally furthers a discriminatory culture, but it is not obvious to other people. There were arguably too many words on the list, and people did not understand the potential benefits. As social psychologists William Cox said, in an interview with Nordell (2021, p. 105): "We can't change people's values, but we can give people knowledge about how they might not be living up to their values. Once you have this information, you can't help but make an effort."

As the initiative by Stanford University was canceled almost immediately, we cannot know what effects it could have had. There is an emphasis in the literature on avoiding mandatory interventions. Dobbin and Kalev (2018), for instance, found that voluntary training is significantly more beneficial than mandatory training. They propose to "give employees a choice of different types of diversity training," (Dobbin & Kalev, 2018, p. 52). Along these lines, it could perhaps be beneficial to suggest or even reward more inclusive language. A lot of people seemed to have felt pressured and/or manipulated to use certain words. In addition, people are susceptible to be persuaded by arguments ((Koeser & Sczesny, 2014, p. 548)). If the initiative had been better explained, and perhaps commercialized, perhaps the results would have been different. The two interventions from Sweden indicated, however, that people become more

accepting of the changes over time. People with negative attitudes toward transgender individuals perceive greater difficulty in using gender-inclusive language (Patev, et. al., 2018). These findings suggest that inclusive language use may be indicative of more positive attitudes.

In conclusion, there is not much evidence on gendered language reducing *implicit* bias. The research is often done on a number of factors, and reducing implicit bias is only one of them. In addition, it is often difficult to distinguish between implicit and explicit bias. Implicit bias is difficult to measure, people are unwilling to admit to them. Implicit bias is an *implicit* phenomenon. Banaji and Greenwald (2013) postulate that "In-group favoritism may be the largest contributing factor to the relative disadvantages experienced by Black Americans and other already disadvantaged groups." But we do not see it, and many of the people taking the IAT tests, two of the creators, Banaji and Greenwald (2013) included, have egalitarian beliefs, but still tests for having, for instance, automatic white preference. They further criticizes Dovidio & Gaertner (2004) for using the term aversive racism to refer to people that test for automatic white preference *racists*. Here is Dovidio and Gaertner (2004):

> Thus, it is important that people, including both whites and blacks, become aware of the existence and impact of aversive racism to understand the different perspectives of members of different racial groups, to facilitate more effective communication and, ultimately, to take appropriate personal, social, and legal action to create a fully egalitarian society. Because aversive racists genuinely endorse egalitarian principles, once aware of their biases, they can help contribute to the solution rather than to the problem of racial tension, conflict, and inequality.

While their main point is valid, and that the problem might lie with people who genuinely do not think of themselves as the problem, the language use is misguided, and puts the participants in an unwanted category. According to (Banaji & Greenwald, 2013) they should instead be called uncomfortable egalitarians:

> It is unwarranted to attach the racist label to the many people that show an automatic white preference on the IAT [and] extremely unlikely to notice that their differential behavior of Whites and Blacks contributes in any way to the disadvantages experienced by Black Americans

As they point out, no one accuses you of discriminating if you give your kidney to a sibling rather than to a stranger. Nepotistic hiring, on the other hand, is legally discrimination.

Banaji and Greenwald (2013) conclude that implicit bias can often be outsmarted, but rarely eradicated. Some of the proposals they suggests are labeling something as a phobia, instead of a stereotype: "The significance of labeling something as a phobia instead of a

stereotype is that the cause of aversion seems to be a property of responding to the dog, rather than a property of the dog," (Banaji & Greenwald, 2013). Banaji and Greenwald (2013) conclude that ""Alas, there has not yet been a convincing, to us, demonstration that interventions of the types investigated in research of the last decade will produce durable changes." Even when there is a reduction in implicit bias, it seems that it is an elastic change, and not a permanent change (Dasgupta, et. al., 2000). It is unclear what to do about this.

4.2 Is Conceptual Engineering an Effective Strategy for Reducing Implicit Bias?

The research question of this thesis is what the prejudice literature conveys about the effectiveness in reducing implicit bias. To address this research question I focused on one approach to CE, which is to de-gender language, as this has been argued by philosophers in the field of CE (Dembroff & Wodak, 2018; Ritchie, 2021). One criticism here could be that the hypothesis should have been the research question. I chose not to do this because the aim is to consider CE as a framework for reducing implicit bias, and I wanted to connect the predominant philosophical discipline of CE with the predominantly psychological literature on implicit bias. By considering the effects of de-gendering language, I hope to have achieved this. In addition, I believe that the research question of this paper deserves attention, in both a systematic review and in diversity training programs like implicit bias training.

The benefits of doing a scoping review is that it can inspire a systematic review. A next step would be to do a systematic review of what the literature conveys on CE's effectiveness on reducing implicit bias. A systematic review could do a better job assessing the methods applied in the different studies mentioned in this thesis. Furthermore, many more aspects of approaches to language change could be included, like changing perceived norms to make people behave according to the norm regardless of what they think about it (Cialdini, et. al., 2006; Nimtz, 2021). More importantly, based on these conclusions, CE should be considered as part of implicit bias training. It has to be done differently and more tactfully than Stanford University's attempt. Snapshot interventions do not seem to have much effect. What is needed is corporations or universities to facilitate training over a number of years (Dobbin & Kalev, 2018,). Based on the findings here, language change could be a part of that. If CE is a *part of* implicit bias training, the language changes become less prominent, and therefore less likely to cause adverse reactions.

What about the author of this paper? I use the terminology predominant in the literature, like *White Americans* and *Black Americans*. As these are adjectives it should not cause essentializing. However, does the capital letters cause more *categorization*? Furthermore, there is a tendency in literature to write *Whites* and *Blacks* (e.g., Dobin & Kalev, 2018). So, even in the prejudice literature there is a tendency to use essentializing nouns (cf. Ritchie, 2021). I had to correct myself on occasion. One could argue, of course, that if there is one area where this should be allowed it has to be where it is being researched. One could also argue that if there is one area where they should be responsible and avoid using essentializing nouns and gendered language it is the prejudice literature.

5. Conclusion

A scoping review was conducted to investigate the potential of using CE to reduce implicit bias. The scope was narrowed to reviewing whether reducing gendered language can reduce implicit bias. The evidence is inconclusive. There are the additional problems of measurement, like IAT, and unproven assumptions, like linguistic relativity. The research question is directly associated with linguistic relativity, and is dependent on this being true. The support for linguistic relativity is mixed, and in much of the research there is a problem of ruling out confounding variables. Similarly, how can researchers on gendered language conclude conclusively that gendered languages, versus gender-neutral languages, is what signifies the gender gap in education, for instance. All of this depends on a form of linguistic relativity. Similarly, how can researchers on gendered language conclude conclusively that gendered languages, versus gender-neutral languages, is what signifies the gender gap in education, for instance. All of this depends on a form of linguistic relativity.

While there is a clear correlation between gendered language and implicit bias, it is unconvincing whether language change interventions can have the desirable effects that philosophers argue. A systematic review on the effects of CE in reducing implicit bias is called for. Furthermore, the evidence of using CE to reduce implicit bias is promising enough to be included in diversity programs, like implicit bias training. It would be of particular interest to study the inclusion of the gender-neutral pronoun “hen” in Norway and Sweden, and perhaps a comparison study between the two, since Sweden introduced it 10 years before Norway.

References

Adam, H. & Galinsky, A. D. (2012). Enclothed cognition. *Journal of Experimental Social Psychology, 48*(4), 918-25.

Adams, G., Biernat, M., Branscombe, N. R., Crandall, C. S. & Wrightsman, L. S. (2008). Beyond Prejudice: Toward a sociocultural psychology of racism and oppression. In G. Adams, M Biernat, N. R. Branscombe, C. S. Crandall & L. S. Wrightsman (Eds.), *Commemorating Brown: The social psychology of racism and discrimination* (pp. 215-46). American Psychological Association. DOI: 10.1037/11681-012

American Psychological Association (2023a). Implicit Attitude. https://dictionary.apa.org/implicit-attitude

American Psychological Association (2023b). Categorization. https://dictionary.apa.org/categorization

American Psychological Association (2023c). Types of Articles Accepted. https://www.apa.org/pubs/journals/bar/article-types

Appiah, K. A. (2018). *The Lies That Bind: Rethinking Identity*. New York: Liveright.

Appiah, K. A. (2022). Engineering Race. Online Seminar, at *Arché Philosophical Research Centre*, University of St. Andrews, March 29th. https://www.youtube.com/watch?v=BpJclAKqT-g

Applebaum, B. (2019). Remediating campus climate: Implicit bias training is not enough. *Studies in Philosophy and Education*, *38*, 129-141

Aromataris, E., & Pearson, A. (2014). The systematic review: an overview. *AJN The American Journal of Nursing*, *114*(3), 53-58.

Ashford, R. D., Brown, A. M., & Curtis, B. (2018). Substance use, recovery, and linguistics: The impact of word choice on explicit and implicit bias. *Drug and alcohol dependence*, *189*, 131-138.

Banaji, M. R. & Greenwald, A. G. (2013). *Blindspot: Hidden Biases of Good People*. Random House Publishing Group.

Barkved, B. I. (2022). All Words Are Equal, But Some Words Are More Equal Than Others: What the scope of conceptual engineering should be.

Bell, J. (2010). *Doing Your Research Project: A guide for first-time researchers in Education, Health and Social Science* (5th ed.). Open University Press.

Bender, A., Beller, S., & Klauer, K. C. (2016). Crossing grammar and biology for gender categorisations: Investigating the gender congruency effect in generic nouns for animates. *Journal of Cognitive Psychology*, *28*(5), 530-558.

Bezrukova, K., Spell, C. S., Perry, J. L., & Jehn, K. A. (2016). A meta-analytical integration of over 40 years of research on diversity training evaluation. *Psychological bulletin*, *142*(11), 1227.

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Bicchieri, C. (2016). *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press.

Billig & Tajfel (1973).

Boroditsky, L. (2001). Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognitive psychology*, *43*(1), 1-22.

Boroditsky, L., Fuhrman, O., & McCormick, K. (2011). Do English and Mandarin speakers think about time differently?. *Cognition*, *118*(1), 123-129.

Boroditsky, L., Schmidt, L. A., & Phillips, W. (2003). Sex, syntax, and semantics. *Language in mind: Advances in the study of language and thought*, *22*, 61-79.

Brownstein, M. (2019). Implicit Bias. *The Stanford Encyclopedia of Philosophy*, E. N. Zalta (Ed.). <https://plato.stanford.edu/archives/fall2019/entries/implicit-bias/>.

Cappelen, H. (2018). *Fixing Language*. Oxford University Press.

Carpenter, T. P., Goedderz, A., & Lai, C. K. (2022). Individual differences in implicit bias can be measured reliably by administering the same Implicit Association Test multiple times. *Personality and Social Psychology Bulletin*.

Casati, R. (2022). Reconceptualizing the Ocean. Online Seminar, at *Arché Philosophical Research Centre*, University of St. Andrews, April 27th. https://www.youtube.com/watch?v=LqrPdJzSl1Y&t=29s.

Chapman, E. N., Kaatz, A., & Carnes, M. (2013). Physicians and implicit bias: how doctors may unwittingly perpetuate health care disparities. *Journal of general internal medicine*, *28*, 1504-1510.

Churchland, P. (2021). Social Conscience: Evolutionary Origins and Brain Mechanisms.

Online Seminar, at *Arché Philosophical Research Centre* at University of St. Andrews and *Department of Philosophy* at University of Zürich, September 14th. https://www.youtube.com/watch?v=Yo-XL6soOvg.

Clark, A. & Chalmers, D. (1998). The Extended Mind. *Analysis, 58*(1), 7-19.

Clark, K. B., & Clark, M. P. (1950). Emotional factors in racial identification and preference in Negro children. *Journal of Negro education*, *19*(3), 341-350.

Corbett, G. G. (1991). *Gender*. Cambridge: Cambridge University Press.

Dasgupta, N., McGhee, D. E., Greenwald, A. G., & Banaji, M. R. (2000). Automatic preference for White Americans: Eliminating the familiarity explanation. *Journal of Experimental Social Psychology*, *36*(3), 316-328.

Davidson, D. (1967). Causal relations. *The Journal of philosophy*, *64*(21), 691-703.

Davis, L., & Reynolds, M. (2018). Gendered language and the educational gender gap. *Economics letters*, *168*, 46-48.

Dembroff, R. & Wodak, D. (2018). He/She/They/Ze. *Ego*, *5*(14), 371-406.

Devos, T., & Banaji, M. R. (2005). American= white?. *Journal of personality and social psychology*, *88*(3), 447.

Dixon-Woods, M., Cavers, D., Agarwal, S., Annandale, E., Arthur, A., Harvey, J., Hsu, R., Katbamna, S., Olsen, R., Smith, L., Rile, R. & Sutton, A. J. (2006). Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC medical research methodology*, *6*, 1-13.

Dobbin, F., & Kalev, A. (2016). Why diversity programs fail. *Harvard Business Review*, *94*(7), 14.

Dobbin, F., & Kalev, A. (2018). Why doesn't diversity training work? The challenge for industry and academia. *Anthropology Now*, *10*(2), 48-55.

Dovidio, J. F., & Gaertner, S. L. (2004). Aversive racism.

Drayton, L. (2023, May 9th). Preparation of Opinion Manuscripts. *Trends in Cognitive Sciences*. https://www.cell.com/trends/cognitive-sciences/authors.

Duflo, E. (2012). Women empowerment and economic development. *Journal of Economic literature*, *50*(4), 1051-1079.

Erdol, T. A. (2019). Practicing gender pedagogy: The case of Egalia. *Eğitimde Nitel Araştırmalar Dergisi*, *7*(4), 1365-1385.

Everett, C. (2013). *Linguistic Relativity: Evidence across languages and cognitive domains.* De Gruyter Mouton. DOI: 10.1515/9783110308143

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I", the "A", and the "T": A logical and psychometric critique of the Implicit Association Test (IAT). *European review of social psychology*, *17*(1), 74-147.

Filippou, P., Mahajan, S., Deal, A., Wallen, E. M., Tan, H. J., Pruthi, R. S., & Smith, A. B. (2019). The presence of gender bias in letters of recommendations written for urology residency applicants. *Urology*, *134*, 56-61.

Fischer, E. (2020). Conceptual control: On the feasibility of conceptual engineering. *Inquiry*, 1-29.

FitzGerald, C., Martin, A., Berner, D., & Hurst, S. (2019). Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: a systematic review. *BMC psychology*, *7*(1), 1-12.

Fodor, J. (1975). *The Language of Thought*. Harvard University Press.

Furley, P., & Goldschmied, N. (2021). Systematic vs. narrative reviews in sport and exercise psychology: Is either approach superior to the other?. *Frontiers in psychology*, *12*, 685082.

Gallagher, Steve (2022). Update on Elimination of Harmful Language Initiative in Stanford's IT Community. *Stanford University*, https://itcommunity.stanford.edu/news/update-elimination-harmful-language-initiative-stanfords-it-community

Gay, V., Santacreu-Vasut, E., & Shoham, A. (2013). The grammatical origins of gender roles. *Berkeley Economic History Laboratory Working Paper*, *3*.

Gay, V., Hicks, D. L., Santacreu-Vasut, E., & Shoham, A. (2018). Decomposing culture: an analysis of gender, language, and labor supply in the household. *Review of Economics of the Household*, *16*, 879-909.

Ghosh, S., & Caliskan, A. (2023). ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. *arXiv preprint arXiv:2305.10510*.

Goff, P. A. (2021). Perspectives on Policing: Phillip Atiba Goff. *Annual Review of Criminology*, *4*, 27-32.

Gojkov-Rajić, A., & Prtljaga, J. (2013). Foreign language learning as a factor of intercultural tolerance. *Procedia-Social and Behavioral Sciences*, *93*, 809-813.

Gonzalez, C. M., Lypson, M. L. & Sukhera, J. (2021). Twelve tips for teaching implicit bias recognition and management. *Medical Teacher, 43*(12), 1368-73.

Grant, M. J., & Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health information & libraries journal*, *26*(2), 91-108.

Greenwald, A. G. & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem and stereotypes. *Psychological review, 102*(1), 4-27.

Greenwald, A. G., Banaji, M. R. & Nosek, B. A. (2023). Project Implicit. https://implicit.harvard.edu/implicit/takeatest.html.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of personality and social psychology*, *85*(2), 197.

Greenwald, A. G. & Lai, C. K. (2020). Implicit social cognition. *Annual review of psychology, 71*, 419-45.

Greenwald, A. G., Smith, C. T., Sriram, N., Bar‐Anan, Y., & Nosek, B. A. (2009). Implicit race attitudes predicted vote in the 2008 US presidential election. *Analyses of Social Issues and Public Policy*, *9*(1), 241-253.

Griffin, D. W., Gonzalez, R., Koehler, D. J. & Gilovich, T. (2012). Judgmental heuristics: A historical overview. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 322–345). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199734689.013.0017

Gustafsson Sendén, M., Bäck, E. A., & Lindqvist, A. (2015). Introducing a gender-neutral pronoun in a natural gender language: the influence of time on attitudes and behavior. *Frontiers in psychology*, *6*, 893.

Gygax, P. M., Elmiger, D., Zufferey, S., Garnham, A., Sczesny, S., Von Stockhausen, L., … & Oakhill, J. (2019). A language index of grammatical gender dimensions to study the impact of grammatical gender on the way we perceive women and men. *Frontiers in Psychology*, *10*, 1-6.

Hart, C. (2018). *Doing a literature review: Releasing the social science research imagination*. Sage Publications.

Haslanger, S. (2000). Gender and race: (What) are they? (What) do we want them to be? *Noûs*, 34.1: 31-55.

Haslanger, S. (2012). *Resisting Reality: Social Constructions and social critique*. Oxford University Press.

Herfeld, C. (2022). Thick concepts in Economics. Online Seminar, at *Arché Philosophical Research Centre* at University of St. Andrews and *Department of Philosophy* at University of Zürich, June 7th.

Hill, E., Tiefenthäler, A., Triebert, C., Jordan, D., Willis, H. & Stein, R. (2022, January 24). How George Floyd Was Killed in Police Custody. *The New York Times*. https://www.nytimes.com/2020/05/31/us/george-floyd-investigation.html.

Hummert, M. L., Garstka, T. A., O'Brien, L. T., Greenwald, A. G., & Mellott, D. S. (2002). Using the implicit association test to measure age differences in implicit social cognitions. *Psychology and aging*, *17*(3), 482.

Isaac, M. G. (2023). Conceptual Engineering Online Series. Hosted by *The Department of Philosophy* at the University of Zurich and *Arché Philosophical Research Centre* at the University of St. Andrews. https://www.philosophie.uzh.ch/de/research/congresses/archive/ceos.html.

Isaac, M. G., Koch, Steffen & Nefdt, Ryan (2022). Conceptual Engineering: A road map to practice. *Philosophy Compass*, 17: 1-15.

Jackson, J. L. (2018). The non-performativity of implicit bias training. *The Radical Teacher*, (112), 46-54.

January, D., & Kako, E. (2007). Re-evaluating evidence for linguistic relativity: Reply to Boroditsky (2001). *Cognition*, *104*(2), 417-426.

Jost, 2015. “Resistance to change: A social psychological perspective.” *Social Research*, 82.3: 607-36.

Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system‐justification and the production of false consciousness. *British journal of social psychology*, *33*(1), 1-27.

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political psychology*, *25*(6), 881-919.

Juergensmeyer, M. (2020, September). The Global Context of European Religious

Neo-Nationalism. In *Religion and Neo-Nationalism in Europe* (pp. 49-60). Nomos Verlagsgesellschaft mbH & Co. KG.

Kahneman, D. (2011). *Thinking, Fast and Slow.* Penguin Books.

Kahneman, D., Sibony, O. & Sunstein, C. R. (2021). *Noise*. HarperCollins Publishers.

Kahneman, D. & Tversky, A. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-31.

Kanheman, D. & Tversky, A. (1979). Prospect theory. *Economicra, 47*, 263-91.

Kahneman, D. & Tversky, A. (1982). Judgment of and by representativeness. In D. Kahneman, P. Slovic and A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Kang, Y., Gray, J. R. & Dovidio, J. F. (2014). The nondiscriminating heart: Lovingkindness meditation training decreases implicit ingroup bias. *Journal of Experimental Psychology: General, 143*(3), 1306-13. DOI: 10.1037/a0034150

Kim, J. Y., & Roberson, L. (2022). I'm biased and so are you. What should organizations do? A review of organizational implicit-bias training programs. *Consulting Psychology Journal*, *74*(1), 19.

Koeser, S., & Sczesny, S. (2014). Promoting gender-fair language: The impact of arguments on language use, attitudes, and cognitions. *Journal of Language and Social Psychology*, *33*(5), 548-560.

Koslow, A. (2022). Meaning change and changing meaning. *Synthese*, 200.2: 1-26.

Kousta, S. T., Vinson, D. P., & Vigliocco, G. (2008). Investigating linguistic relativity through bilingualism: the case of grammatical gender. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 843.

Krieger, L. H., & Fiske, S. T. (2006). Behavioral realism in employment discrimination law: Implicit bias and disparate treatment. *California Law Review*, *94*(4), 997-1062.

Legault, L., Gutsell, J. N., & Inzlicht, M. (2011). Ironic effects of antiprejudice messages: How motivational interventions can reduce (but also increase) prejudice. *Psychological Science*, *22*(12), 1472-1477.

Li, J. (2022). Relationship Between Language and Thought: Linguistic Determinism, Independence, or Interaction? *Journal of Contemporary Educational Research, 6*(5), 32-7.

Locke, J. (1690/2008). *An Essay Concerning Human Understanding* (P. Phemister, Ed.). Oxford University Press.

Löhr, G. (2022). Linguistic Interventions and the Ethics of Conceptual Disruption. *Ethical Theory and Moral Practice*, 1-15.

Lucy, J. A., & Gaskins, S. (2003). Interaction of language type and referent type in the development of nonverbal classification preferences. *Language in mind: Advances in the study of language and thought*, 465-492.

Machery, E. (2021). A new challenge to conceptual engineering. *Inquiry*, 1-24.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Jetten, J. (1994). Out of mind but back in sight: Stereotypes on the rebound. *Journal of personality and social psychology*, *67*(5), 808.

Magee, R. V. (2016). The way of ColorInsight: Understanding race and law effectively through mindfulness-based ColorInsight practices. *Georgetown Journal of Law & Modern Critical Race Perspective*, *8*, 251-304.

Mavisakalyan, A. (2015). Gender in language and gender in employment. *Oxford Development Studies*, *43*(4), 403-424.

Miyazono, K., & Bortolotti, L. (2021). *Philosophy of Psychology: An Introduction*. John Wiley & Sons.

Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC medical research methodology*, *18*, 1-7.

Murdock, N. L., & Forsyth, D. R. (1985). IS GENDER‐BIASED LANGUAGE SEXIST? A PERCEPTUAL APPROACH. *Psychology of Women Quarterly*, *9*(1), 39-49.

Neufeld, E. (2022). Psychological essentialism and the structure of concepts. *Philosophy Compass*, *17*(5), e12823.

Noon, M. (2018). Pointless diversity training: Unconscious bias, new racism and agency. *Work, employment and society*, *32*(1), 198-209.

Nordell, J. (2021). *The End of Bias: Can We Change Our Minds?* London: Granta Books.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, research, and practice*, *6*(1), 101.

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., ... & Greenwald, A. G. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, *106*(26), 10593-10597.

Nimtz, C. (2021). Engineering concepts by engineering social norms: solving the implementation challenge. *Inquiry,* 1-28. DOI: 10.1080/0020174X.2021.1956368

Okonofua, J. A., Walton, G. M., & Eberhardt, J. L. (2016). A vicious cycle: A social–psychological account of extreme racial disparities in school discipline. *Perspectives on Psychological Science*, *11*(3), 381-398.

Olsson, A., & Phelps, E. A. (2004). Learned fear of "unseen" faces after Pavlovian, observational, and instructed fear. *Psychological science*, *15*(12), 822-828.

Pankey, T., Alexander, L., Carnes, M., Kaatz, A., Kolemainen, C., Filut, A., ... & Stahr, A. (2018). Breaking the bias-habit: a workshop to help internal medicine residents reduce the impact of implicit bias. *Understanding Interventions*, *9*(2).

Patev, A. J., Dunn, C. E., Hood, K. B., & Barber, J. M. (2019). College students' perceptions of gender-inclusive language use predict attitudes toward transgender and gender nonconforming individuals. *Journal of Language and Social Psychology*, *38*(3), 329-352.

Piaget, J. & Inhelder, B. (1948/1967). *The Child's Conception of Space*. Norton.

Pinder, M. (2021). Conceptual engineering, metasemantic externalism and speaker-meaning. *Mind*, *130*(517), 141-163.

Pinder, M. (2022). What ought a fruitful explicatum to be? *Erkenntnis, 87*(2), 913-32.

Pollock, J. (2019). Conceptual engineering and semantic deference. *Studia Philosophica Estonica*, 81-98.

Regier, P. S., & Redish, A. D. (2015). Contingency management and deliberative decision-making processes. *Frontiers in Psychiatry*, *6*, 76.

Reines, M. F. & Prinz, J. (2009). Reviving Whorf: The return of linguistic relativity. *Philosophy Compass, 4*(6), 1022-32.

Richardson, L. S., & Goff, P. (2013). Implicit racial bias in public defender triage. *Yale Law Journal*, *122*, 13-24.

Ridley, D. (2012). *The literature review: A step-by-step guide for students* (2nd ed.). Sage Publications.

Rosenthal, R. (1973). The Pygmalion Effect Lives. *Psychology Today*.

Rosenthal, R., & Babad, E. Y. (1985). Pygmalion in the gymnasium. *Educational leadership*, *43*(1), 36-39.

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The urban review*, *3*(1), 16-20.

Rudestam, K. E., & Newton, R. R. (2014). *Surviving your dissertation: A comprehensive guide to content and process* (3rd ed.). Sage Publications.

Rudman, L. A., & Heppen, J. B. (2003). Implicit romantic fantasies and women's interest in personal power: A glass slipper effect?. *Personality and Social Psychology Bulletin*, *29*(11), 1357-1370.

Samuel, S., Cole, G., & Eacott, M. J. (2019). Grammatical gender and linguistic relativity: A systematic review. *Psychonomic bulletin & review*, *26*, 1767-1786.

Santacreu-Vasut, E., Shoham, A., & Gay, V. (2013). Do female/male distinctions in language matter? Evidence from gender political quotas. *Applied Economics Letters*, *20*(5), 495-498.

Schwitzgebel, E. (2010). Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly, 91*(4), 531-53.

Shutts, K., Kenward, B., Falk, H., Ivegran, A., & Fawcett, C. (2017). Early preschool environments and gender: Effects of gender pedagogy in Sweden. *Journal of experimental child psychology*, *162*, 1-17.

Stanovich, K. E. (1999). *Who is rational?: Studies of individual differences in reasoning*. Psychology Press.

Stanovich, K. E., & West, R. F. (2000). 24. Individual Differences in Reasoning: Implications for the Rationality Debate?. *Behavioural and Brain Science*, *23*(5), 665-726.

Stock, K. (2022). The importance of referring to human sex in language. *Law & Contemp. Probs.*, *85*, 25.

Sunstein, C. (2021). The Right Not to Be Manipulated. Online Seminar, Online Seminar at *Arché Philosophical Research Centre* at University of St. Andrews and *Department of Philosophy* at University of Zürich, September 21st. https://www.youtube.com/watch?v=_No5iALXd60&t=214s.

Tanesini, A. (2022). Engineering Autobiographical and Collective Memories. at *Arché*

Philosophical Research Centre at University of St. Andrews and *Department of Philosophy* at University of Züric, May 10th. https://www.youtube.com/watch?v=L_tonWKOAk4.

Thaler, R. H. (2016). *Misbehaving: The Making of Behavioral Economics.* W.W. Norton & Company.

Tremain, S. L. (2021). Engineering (the Apparatus) of Disability. Online Seminar at *Arché Philosophical Research Centre* at University of St. Andrews and *Department of Philosophy* at University of Züric, October 19th. https://www.youtube.com/watch?v=1F6W2QHhuqc&t=5s.

Vaughan, G. M., Tajfel, H., & Williams, J. (1981). Bias in reward allocation in an intergroup and an interpersonal context. *Social Psychology Quarterly, 44*(1): 37-42.

Vitevitch, M. S., Sereno, J., Jongman, A., & Goldstein, R. (2013). Speaker sex influences processing of grammatical gender. *PloS one*, *8*(11), e79701.

Von Neumann, J. & Morgenstern, O. (1944/2007). *Theory of Games and Economic Behavior*. Princeton University Press.

Xiao, H., Strickland, B., & Peperkamp, S. (2023). How fair is gender-fair language? Insights from gender ratio estimations in French. *Journal of Language and Social Psychology*, *42*(1), 82-106.