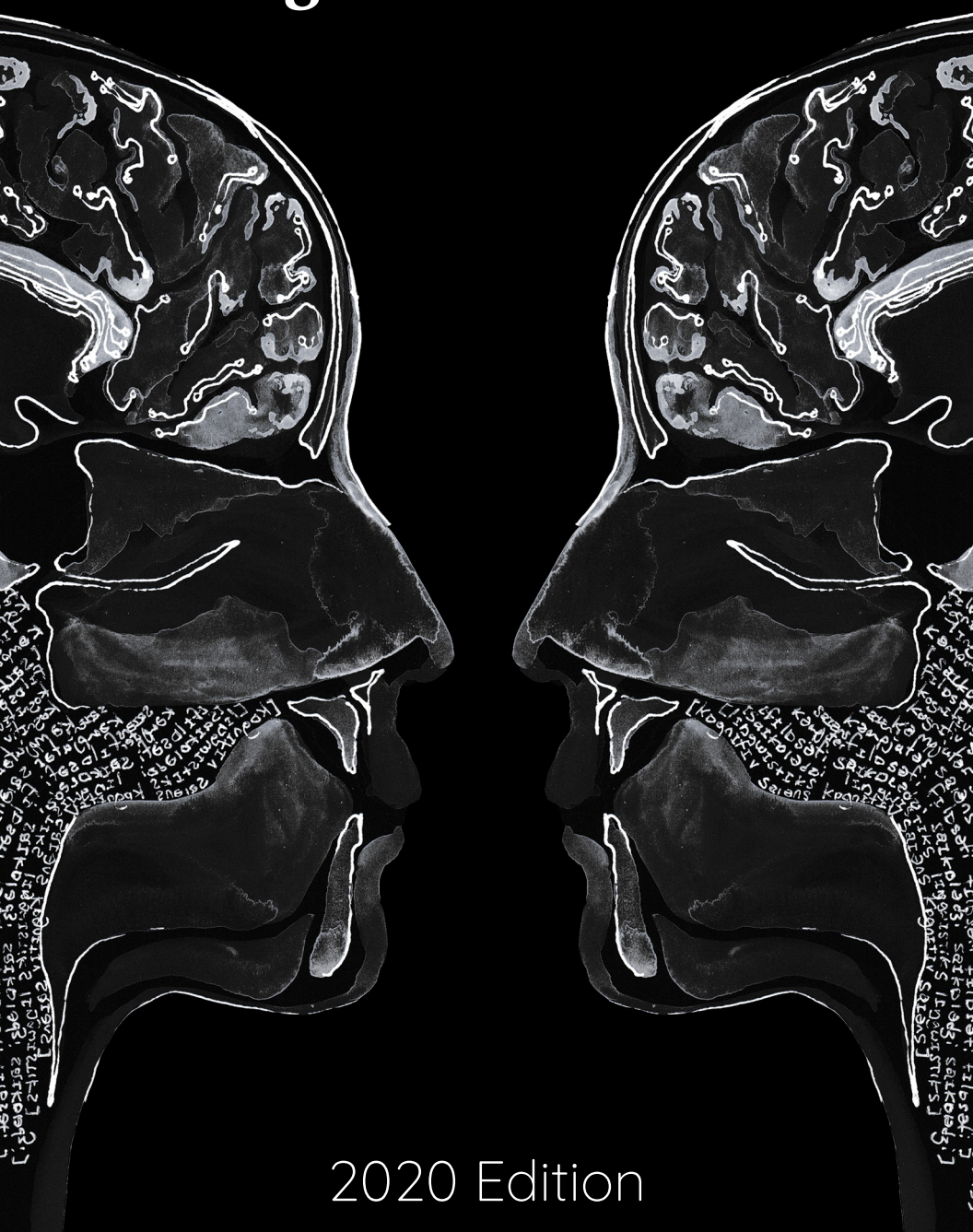
Canadian Undergraduate Journal of **Cognitive Science**



2020 Edition

Canadian Undergraduate Journal of **Cognitive Science** 2020

By the Cognitive Science Student Society of
Simon Fraser University

Simon Fraser University
Burnaby, British Columbia

ISSN 1913-0651 (Print)
ISSN 1499-7487 (Online)

Canadian Undergraduate Journal of **Cognitive Science**

2020 Edition

The Canadian Undergraduate Journal of Cognitive Science (CUJCS) is an academic journal aimed at providing undergraduate students in cognitive science and related fields an opportunity to publish and showcase their work. We seek to foster academic interest and creativity and provide a forum for connections and the exchanging of ideas. As a publication, CUJCS provides a unique reference for students, improves the contact and exchange of ideas between students and cognitive scientists alike, and illustrates the interdisciplinary work that is the hallmark of cognitive science.

Territorial Acknowledgement
This edition of CUJCS has been compiled and published on the unceded Traditional Coast Salish Lands, including the Sḵwx̱wú7mesh Úxwumixw (Squamish), səl̓ílw̓ətaʔɬ (Tsleil-Waututh), xʷməθkʷəy̓əm (Musqueam) and kʷikʷəƛ̓əm (Kwikwetlem) Nations.

cognitive science student society
simon fraser university

simon fraser student society

Table of Contents

Articles

Art Features

The 2020 edition of the Canadian Undergraduate Journal of Cognitive Science has been a long time coming. What started out as a sparkle in the eye of a few cognitive science students at Simon Fraser University back in 2002 has grown to become a journal featuring diverse works from around the world, showcasing the multidisciplinary nature of cognitive science through thought provoking pieces of both research and art. It is that same spark that lit the fire to bring together the 2020 edition before you now.

The goal of the journal has always been to provide a venue for undergraduates in the cognitive sciences to share their work and demonstrate their skills. CUJCS is a place of learning, giving students experience in the processes involved in the publication of an academic journal. We are pleased to say that this year has been a greater success than any previous, with dozens of submissions coming in from not just Canadian universities, but also from undergraduate scholars across the globe. The board is proud to showcase these diverse works and of the experiences given to everyone along the way.

In order to decrease the barriers of access to cognitive science, we have worked to provide unique ways for students to explore the discipline. Reflecting on what cognitive science meant to them, artists too, submitted their work to the journal, five pieces of which are displayed within.

The production of the 2020 edition of the Canadian Undergraduate Journal of Cognitive Science has been a tour de force from outset to publication. It would not have been possible without the contributions and efforts of everyone involved. A special thanks goes out to Dr. Tupper, Dr. Grant, Dr. Sigal, and Dr. Blair whose expertise, guidance, and facilities were paramount to our success. To each and every author who submitted a paper and to all the artists who created wonderful pieces, this journal would not be possible without the fruits of your labour. To the rest of the managerial and editorial board, without whom the journal would not be where it is today, thank you. And finally, to you, the reader. Whether you picked up this journal on a whim or are reading for academic pursuits, your engagement is what drives each and every one of us to produce the highest quality work.

Sincerely,

Rollin Poe
Director, 2020

Kiezdeutsch: Perspectives on Language Ideologies and Variation in German Society

Jeremy Li
Simon Fraser University

In recent decades, a new variety of German called Kiezdeutsch or **neighbourhood German** arose as Germany experienced waves of immigration from neighbouring countries since the late 1950s. Even today, researchers and non-linguists remain divided on the definition of this linguistic phenomenon. Sociolinguistic research conducted in this area uncover firmly rooted language ideologies in Germany—ones which challenge the reality of linguistic variation and change of a standard spoken norm. From this, we gain a general understanding of the subtle nature of societal forces that shape our attitudes towards non-standard linguistic forms and their speakers. As Kiezdeutsch is a relatively new occurrence, some sociological areas of research remain unexplored; however, such areas can benefit from more well-rounded, collaborative approaches in future studies.

Introduction

A **dialect**, as described by linguists and non-linguists, is a variety of spoken code typically considered subordinate to a standard speech norm, or **language**. Despite this view, many dialects around the world, such as Swiss German, are used extensively in everyday contexts and enjoy relatively favourable status among its peers. Moreover, what is considered a dialect, as opposed to a speech style or "deviation", varies from context to context. In Germany, a dialect, or *Dialekt*, is considered one of many "true" varieties spoken by ethnic Germans living in demarcated regions of Germany (Wardhaugh & Fuller, 2015); these regional varieties are ingrained in speakers' identities and throughout Germany's extensive regional histories. With the emergence of immigrant varieties of German in recent decades, discussion of whether these are "legitimate" or "corrupted" forms remain contentious (Stevenson, 2017).

In this paper, I explore the literature on *Kiezdeutsch*, a so-called "street" German that holds a polarising status in the German language community. To begin, I look at its history—from its conception to today—and its users. As well, I briefly describe several salient linguistic features of Kiezdeutsch, providing examples of its phonology, syntax, and lexicon.

Next, I examine literature on language attitudes and ideologies in Germany. In particular, I explore Stevenson's (1997) ideas of the standard German language as a "national asset" and a benchmark for "Germanness". To better understand these perspectives, I first clarify how *Hochdeutsch,* or **Standard German** became the standard speech norm. Then, I seek to establish relationships, if any, between language attitudes and its portrayal of Kiezdeutsch, considering cultural sensitivities to language change and usage. To support this, I provide a parallel using English's influence on the German language, demonstrating that language attitudes are not confined to a particular variety, but to non-standard variation, in general.

Finally, I summarise two main perspectives in recent literature—that of **style** versus **variety**—which researchers use to define Kiezdeutsch as a linguistic phenomenon. Their research addresses sociolinguistic theories as to why people use Kiezdeutsch, focusing on factors like identity, ethnicity, and social contexts. I accompany this with a perception study conducted by Freywald, Mayr, Özçelik, & Wiese (2011), in hopes of using recorded quantitative and qualitative data to support their findings. I conclude with a review on the literature. In particular, I wish to convey to the reader that defining phenomena like Kiezdeutsch is not a simple task.

A History of Kiezdeutsch and its Users

Since the late 1950s, Germany experienced an unprecedented wave of immigration from neighbouring countries such as Italy, Spain, and Turkey due to labour shortages and the immediate need to fill them. Spurred by necessity, *Gastarbeiterdeutsch* or **guest-worker German** arose as a variety of German used by migrant workers to communicate, utilising linguistic elements of German and the workers' respective languages. Today, researchers are interested in what guest-worker German has evolved into, namely *Kiezdeutsch* or **neighbourhood German,** found especially in urban, multiethnic neighbourhoods. It can be heard being used by the subsequent generations of guest workers,

typically of Turkish, Arabic, and Romani background (Deppermann, 2008). Socialised in Germany, their linguistic repertoires include both standard and non-standard varieties of German as well as other languages (Stevenson, 2017). Researchers characterise Kiezdeutsch as a **youth variety**, as its predominant users—or as Wiese (2012) puts it, "innovators"—are adolescents and young adults.

Stevenson (1997) notes that attempts to explain its earlier form *Gastarbeiterdeutsch* revolved around three popular hypotheses: (1) that it is a result of **interference** or **transfer** from its speakers' native varieties; (2) that it is a **pidgin**, or contact language between groups; or (3), a popular narrative, especially in German-speaking spheres, that it is "foreigner talk", a broken form of German. These theories were often used to reinforce contemptuous public attitudes towards its immigrant speakers and dismiss their linguistic competence. Interestingly, Stevenson points out that speakers of *Gastarbeiterdeutsch* displayed striking similarities in the way they used it despite their varying native languages, suggesting this may be an indicator for something systematic as opposed to random.

In following decades, German linguistic researchers such as Wiese (2012) and Auer (2013) have contributed large amounts of literature on the subject, seeking to elaborate on or debunk these hypotheses. Despite their efforts, this negative, often discriminatory view of "deviant" language use and its speakers persists in the public mind of today's German-speaking societies (Freywald et al., 2011). In fact, popular references to Kiezdeutsch include: *Türkendeutsch* (Turkish German), *Ghettodeutsch* (ghetto German), or *Kanak-sprak* (lit. *foreigner speech*), a comedic stylization of the ethnic slur *Kanake* and *Sprache* (Auer, 2013; Deppermann, 2008; Wiese & Tanış-Polat, 2016). A push in academia towards the more accurate "Kiezdeutsch" to reflect its neighbourhood contexts has taken place in recent literature; however, potentially offensive terms like *Kanak-sprak* remain common.

Linguistic Features of Kiezdeutsch

Before exploring the literature on Kiezdeutsch, readers should get an idea of what Kiezdeutsch entails, linguistically. Generally, its linguistic features are easily discernable to native German speakers when compared to Standard German, especially on the speech level (Preseau, 2018). For example, the realisation of the post-alveolar [ʃ] instead of the standard palatal fricative [ç] is heard in common words

such as the first-person pronoun "*ich*" (i.e. [iʃ] instead of [iç]). In addition, reductions in standard morphology are common, such as omitting morphological inflections demonstrating grammatical case, number, and gender. Locative grammatical constructions such as "*Ich geh' U-bahn*" instead of "*Ich gehe **in die** U-bahn*" (I'm going **to the** subway) are typical, and often seen printed on tourist souvenirs in cities like Berlin (Wiese, 2012, p. 54).[1]† As well, loanwords, especially of Turkish and Arabic origins, are often inserted into utterances, such as *wallah* (I swear) or *yalla* (Let's go!).

Language Attitudes and Ideologies in Germany

To understand language attitudes towards Kiezdeutsch, we must understand its relationship to the standard variety, *Hochdeutsch*, or **Standard German**, as well as regional German varieties. Emerging from a variety used in 17th century German chancelleries, *Hochdeutsch* was the language of the educated, influential, and socially exclusive middle and upper classes (Wiese, 2012). As Wiese (2012) notes, the centuries long (and mistaken) demarcation of this official variety as a "high" language simultaneously relegated varieties such as regional dialects to a low status and cemented its own position as the prestigious standard. Today, *Hochdeutsch* remains the standard spoken norm in Germany, coexisting with regional dialects such as *Plattdeutsch* (Low German varieties in northern Germany) and *Hochdeutsch*[2] (Central and Upper German varieties, such as Bavarian German, in central and southern Germany) (Stevenson, 1997). As local traditions have endured over centuries in their respective regions, regional dialects are one of the many cultural manifestations that persist as distinct and seperate, yet integral parts of the German identity (Barbour, 2000; Stevenson, 1997).

As such, Lippi-Green (2012, in Wardhaugh and Fuller, 2014) attests to the idea of a standard language being an objective and clearly defined variety as a "myth". Stevenson (1997) writes:

> "Far from being a naturally occurring primordial phenomenon, it is always the result of relatively recent and deliberate intervention in the "natural" development of the language." (p. 8)

1 In-text quotations and parentheses marked with a dagger (†) are translations from German to English by me.

2 Not to be confused with Standard Hochdeutsch, but associated with the highland geography of its speaker areas

In other words, the definition of a standard language is not a "natural" occurrence, but rather a subjectively motivated one. As evidenced by its historical development and ascent to the prestigious norm, Standard German is no different. Regardless, the perception, or the **standard language ideology** of Standard German as the prestigious, correct, and grammatically "better" form is firmly rooted within its speakers (Wiese, 2012).

According to Stevenson (1997), the German language is regarded by its speakers—in this case, German nationals—to be two things: (1) a "national asset"; and (2) a benchmark for "Germanness". It is something to be protected from deviations and foreign influences. Consequently, "purist" associations for *Sprachpflege*, or "language care", exist for this reason (Preseau, 2018). Because of these language ideologies, the notion that Standard German is under constant threat is pervasive in German society (Stevenson, 1997). As such, public opinions on its state consistently revolve around a common narrative: that *Sprachverfall*, or "language decay", is happening due to foreign influences.

The notion of *Sprachverfall*, however, is not exclusive to Kiezdeutsch. For example, a parallel can be observed with the pervasiveness of English in Standard German. Since the 17th century to today, English lexical borrowings and their usage have grown exponentially due to increasing contact with and consumption of English media in today's globalised German contexts (e.g. business, politics, academia, music), in which English is often a common language, or **lingua franca** (Hilgendorf, 2007). This phenomenon has led to alarm over a perceived **creolisation** of "traditional" German (Barbour, 2005). Contrary to such concerns—that English is "attacking" or "taking over the German language" (p. 159)—Barbour (2005) argues that English's influence on Standard German is confined to the lexical domain, one where borrowings do not always equate or exist in English (e.g. *Smoking* 'tuxedo', *Oldtimer* 'vintage car', *Handy* 'cellphone'). Despite this, organisations such as the *Verein Deutsche Sprache*, or "German Language Association", continue to perpetuate a narrative of protecting the purity of the German language, one which Barbour (2005) sees as politically motivated versus linguistically. Despite the overt prestige and ubiquity of English in today's global contexts, it is not immune to negative reactions; thus, it is not surprising that Kiezdeutsch would attract similar responses.

In the same vein, language change in Standard German does not only originate from without its speaker spheres, but also from within. Orthographic language reforms (*Reformen der deutschen Rechtschreibungen*) made within recent years, Stevenson (1997) explains, were "met with suspicion… and at worst with outright hostility" (p. 150). Public discourse over these reforms at the time were common and politically charged. The *Duden*, Germany's most well-known reference dictionary, inadvertently plays a role in legitimising backlash, as it holds a de-facto institutional position in society for all matters concerning the German language form. Weinrich (1976, in Stevenson, 1997) states:

> "*In Deutschland verkörpert der Duden die sprachliche Authorität schlechthin.*" (p. 148)
> ("In Germany, the Duden embodies *the* linguistic authority, without exception.")†

As illustration for the power of the *Duden*, Wiese (2012) provides two examples taken from German newspaper articles regarding Kiezdeutsch:

> "'*Kanak Sprak' ignoriert den Duden, und auf eine Notzucht mehr oder weniger an der Grammatik kommt es ihr ebenfalls nicht an.*"
> —Berliner Zeitung, 1999 (p. 143)

> ("'Kanak-Sprak' ignores the *Duden*, and at the same time, a violation/rape of the grammar more or less does not occur to it")†

> "*ein eigenartiges nicht Duden-kompatibles Gossen-Stakkato*"
> — Berliner Morgenpost, 2001 (p. 143)

> ("a peculiar, non-*Duden* compatible gutter-splattering")†

With respect to English, Barbour (2005) asserts the Duden's *Fremdwörterbücher* publications, or "dictionaries of foreign words", also reinforce the idea of "*nicht-germanisch*" or "non-Germanic" designations to English loanwords, which resonate with the public:

> "The pedagogical of informational value of such dictionaries is unclear, but they seem to accord with the linguistic purism of a certain section of society…" (p.154)

In addition to perceived threats and linguistic "violations", a conflation of Kiezdeutsch with "poverty", "antisocial behaviour", and

"language incompetence" is often heard in public discourse. Stevenson (1997) notes, however, that public attitudes regarding language use as an extension of social status and behavior is not a new phenomenon. For example, sociodemographic data of Berlin, Germany expresses a strong correlation between poverty, high migrant concentrations, and economically depressed neighbourhoods, for which Wiese (2012) notes, aligns with neighbourhoods where Kiezdeutsch is spoken. This serves to strengthen a "homegrown" belief that Kiezdeutsch translates to undesirable socioeconomic factors such as unemployment, welfare aid dependence (known in Germany as *Hartz-IV*), and low social prestige (Wiese, 2012). Moreover, Cameron (1995, in Stevenson, 1997) coined the term "verbal hygiene", a response to language cleansing programs "based on the belief that 'ignorance or defiance of grammatical rules is equated with anti-social or criminal behavior'" (p. 168). Although these perspectives may be extreme, they indicate an overt sensitivity regarding language matters—and possibly, a discriminatory undercurrent embedded in language discourse.

As demonstrated, the reactions regarding language change and variation attest to three phenomena: (1) the grip of **hegemonic**, or dominant language ideologies in German society; (2) the power of seemingly benign material like dictionaries which serve to preserve it; and (3) the belief that non-standard, deviant language use is directly connected to disadvantageous socioeconomic factors. Considering this, we begin to understand why Kiezdeutsch has the reputation of "broken German". However, its reputation is not distinguished by its linguistic properties; rather, it is a product of society.

A Discussion of Style Versus Variety

Even in current literature on Kiezdeutsch, researchers remain divided on its classification. On the one hand, researchers like Deppermann (2008), Auer (2013), and Dorleijn & Nortier (2013) characterise Kiezdeutsch as a youth speech "**style**", a multi-functional tool for **self-positioning**; it is a means to signal identity, establish solidarity, or distance oneself from another individual or group[3]. On the other hand, Wiese et al. (2014) and Freywald et al. (2011) take the view that Kiezdeutsch is not merely a speech style. Rather, it is an elaborated German youth dialect which is systematic in its grammar, vocabulary,

3 There is an implication that Kiezdeutsch is something which "augments" a base repertoire for signaling purposes

and social functions. Wiese (2012) calls it a "*Turbo-Dialekt*", or "turbo dialect", that is "more dynamic and more open towards linguistic innovation" (p. 46).†

According to Preseau (2018), using Kiezdeutsch portrays a "tough, ghetto" identity, which carries **covert prestige**, or high status within its speech communities; it possesses an "anti-hegemonic" function expressed through the creation of solidarity within marginalised speaker groups. In other words, it serves as an opposition to integration into a traditional and dominant German society, one which Deppermann (2008) claims is "discriminating and hostile" towards those with migrant backgrounds. Interestingly, Kiezdeutsch has a simultaneous, dual function—not only is it an opposition to traditional societal norms, but a "[refusal] to continue their parents' way of life" (p. 325). This suggests that speakers of Kiezdeutsch signal membership within a third group, one that distinguishes them from the mainstream German culture as well as the one of their parents.

Moreover, Deppermann (2008) explores other functions of Kiezdeutsch—mainly as a **secondary ethnolect**, a comedic instrument or "fun" code that is added into conversations. He models this perspective around earlier comedic, often stereotypical portrayals of Kiezdeutsch in media. (examples found in Deppermann, 2008 p. 331), and its subsequent usage in colloquial conversation among German youth. Interestingly, Hill (1995, in Deppermann, 2008) attributes usages of *Kanak-sprak* with a "double **indexicality**", a simultaneous "pointing" in two directions: one towards creating a funny mood; the other, a "potentially racist and prejudiced out-group identity, that is confirmed and reproduced by the humorous practice (p. 348)". The reason for the latter, Hill suggests, is a "symbolic revenge" as a coping mechanism for "feelings of inferiority" (p. 350).

As well, Rampton (1995; 1998, in Deppermann, 2008) suggests **language crossing** as another function, a way for individuals to participate in or mock—and consequently, distance themselves from—speech communities to which Kiezdeutsch belongs. In the same vein, Auer (2013) brings forth the implication in recent studies of a "de-ethnicising" of Kiezdeutsch, that a perceptual shift from "foreigner talk" to "urban youth style" has taken place in recent decades; he notes its speakers "are no longer only migrant youth, but monolingual German youth, as well (p. 19)".†

On the other hand, the latter group of researchers see Kiezdeutsch as a **dialect** of Standard German; precisely, it is a **multiethnolectal variety**, the result of developments in the standard speech norms in urban, multilingual settings for social participation by youth speakers of various ethnicities. (Preseau, 2018; Wiese et al., 2012, 2014). Freywald et al. (2011) add:

> "A variety should display linguistic features that support a characteristic way of speaking. Seen from an ethnographic perspective, these features should be recognised by its speakers and by other members of the larger community and mark it as distinctive."(p.49)

For instance, their research identifies grammatical and lexical innovations—such as new word positions or additional functions for existing words—that distinguish it from both Standard German and other regional German dialects.[4] As previously mentioned, Stevenson (1997; 2017) found that despite their varying linguistic backgrounds, speakers of its earlier form, *Gastarbeiterdeutsch*, produced similar varieties; he suggests later (specifically in his research on the multilingual context of Berlin, Germany) that the acquisition process of its contemporary form, Kiezdeutsch, is systematic in its development: speakers and non-speakers recognise it as distinctive, and not "simply ungrammatical" or random.

In particular, Wiese at al. (2014) argue that Kiezdeutsch is a "young form of German", an example of ongoing language change emerging from multiethnic, urban areas. In fact, they categorise Kiezdeutsch as a distinct German youth *Dialekt*, whereby speakers are not limited to migrant youth, but extend to German natives, whether mono- or multiethnic. More importantly, Wiese (2012) asserts Kiezdeutsch is not a linguistic compromise to Standard German, but rather used in tandem, dependent on the context of its interlocutors. When approached on the streets of Berlin, Wiese (2012) noted that groups of speakers immediately **code-switched** to Standard German in the presence of adult strangers, concluding that speakers were conscious and selective with whom they spoke—taking into account age, relationship, and peer "in-group" status. As one speaker explains: "I cannot speak to my

4 For a comprehensive analysis of one of the syntactic innovations of Kiezdeutsch, see De Velde's *Temporal adverbs in the kiezdeutsch left periphery: Combining late merge with deaccentuation for v3* (2017)

father like that. It would be so disrespectful." (in Wiese, 2012, p. 116)†

The **dialect** distinction noted earlier, however, is not attributed to the traditional definition of "a [regional] variety under the roof of a standard variety", or **regiolect** (p. 277). Rather, it is a **sociolect**, a dialect of a particular social group—or in this case, a multiethnic, multilingual speech community not confined to a particular space. Both regio- and sociolects, as Wiese (2012) claims, belong under the general term *Dialekt*, the former denoting variation along a horizontal plane (from region to region) and the latter on a vertical plane (from social class to social class).

Auer (2013) is not convinced. Unlike regional varieties—which are historically closely related to Standard German and have unique regional features—Kiezdeutsch is an aggregate of simplified Standard German with other linguistic material (pp. 36-38). Regarding social contexts, such as today's urban centers, he asserts their quick-changing, dynamic nature cannot foster the emergence, much less the solidification of a variety. To do so requires time in stable contexts. He, and other researchers (see Clarke, 1967 in Hinnencamp, 2005), support the "bricolage argument", that variations like Kiezdeutsch are hybrid constructions of readily available linguistic resources to create social style. As Auer (2013) points out, adolescent speakers are constantly constructing new linguistic forms from the wealth of "semiotic resources", or meaningful symbolic material, available in these rapidly changing contexts—collecting, forming, and ultimately, discarding them.[5]

Interestingly, Wiese (2012) agrees with his assessment, but for different reasons. She argues that Kiezdeutsch exploits these dynamic contexts via its users contributing their various multilingual competencies to the development of new forms. With an openness to linguistic experimentation, speakers try new utterances, knowing that "there is not only one way to express something" (p. 46).†

A Qualitative and Quantitative Study on Kiezdeutsch

To support the literature, we look at a perception study conducted by Freywald et al. (2011), where researchers asked mono- and multiethnic German adolescents from mono- and multiethnic neighbourhoods to judge the acceptability of sentences among Standard German, Kiez-

5 According to the *Glossary of Multimodal Terms*, **semiotic resources** is defined as "a means for meaning making" i.e. symbolic material for constructing meaning.

deutsch, and false samples. In order to elicit more accurate responses, researchers asked the participants to base their judgement on whether their friends would say these sentences. As Freywald et al. (2011) explain:

> "This was done to diminish the effect of explicit, prescriptive notions speakers might have, which is particularly important in the case of a low-status variety, where speakers tend to have a high level of '**linguistic insecurity**', that is, where they consider the form they use themselves as the incorrect form if it deviates from the standard." (p. 54-55)

Their findings highlighted some general patterns: (1) compared to false sentences, which were almost universally rejected, Kiezdeutsch samples were accepted between 25% to 59% of the time; (2) participants that lived in multiethnic neighbourhoods accepted Kiezdeutsch samples at more than twice the frequency than those from monoethnic neighbourhoods; and (3) unlike participants from monoethnic neighbourhoods, acceptance rates for Kiezdeutsch were nearly identical in multiethnic neighbourhoods, regardless of the participants' ethnic background. These results reflect not only the relationship of Kiezdeutsch to its multiethnic contexts, they counter the narrative that Kiezdeutsch is simply "broken German".

In qualitative assessments of the same participants, Freywald et al. (2011) found that mono- and multiethnic Germans based their selections on different criteria regarding why they chose to accept or reject a sentence. The former group distanced themselves from Kiezdeutsch, commenting on its "non-German, foreigner" qualities. In other words, a dichotomy between "we" and "they" affected their choices. Conversely, the latter noted that Kiezdeutsch was a language within their periphery; in other words, the variety was often used at school, at home, or within their circle of friends. In essence, the degree of exposure to Kiezdeutsch within their surroundings appeared to influence acceptability rates. Moreover, the findings align with some of the language attitudes encountered so far regarding Kiezdeutsch—namely, an antagonistic one which reinforces a sense of "other".

Reflections on the Literature

On the one hand, there is merit to arguments researchers make regarding the legitimacy of Kiezdeutsch as a dialect; however, they

are not without fault. As observed, the perception study conducted by Freywald et al. (2011) statistically demonstrates that German speakers recognise a distinction between Kiezdeutsch and false sentences; this statistical difference is not a coincidence. However, an explanation beyond solely speaker and listener intuitions must be explored more rigorously. With that being said, the literature arguing for **variety** relies on syntactic "innovations"—many which appear interpretive—as sufficient reasoning. This is not enough, for linguistic structure alone does not define a variety. Sociological factors such as class, age, and gender, and usage among these categories must be considered, as there appears a correlation between these factors and the legitimacy of a variety.

As Wiese (2012), Barbour (2000) and Stevenson (1997) mention, the traditional notion of *Dialekt* recognises regional varieties—considered integral parts of a unified German identity—while ignoring social ones. This is surprising, considering Standard German is the product of a historical upper-class sociolect. Wiese calls for a more encompassing definition that includes sociolects like Kiezdeutsch. Regarding Standard German, its recognition as the standard is not only historically, culturally, and societally defined, but institutionally reinforced and publicly supported. This implies that to recognise Kiezdeutsch would require a shift in public perception, which many see as an attack on German culture and identity. This is not a straightforward task. However, no feasible solutions have been made so far on how to bring about this shift. Despite this, Stevenson's (1997; 2017) **ethnographic** research on German speaking populations provides a detailed, informative starting point for understanding Kiezdeutsch's place in society, highlighting key sociolinguistic issues from German perspectives, and creating discussion for readers interested in language matters.

On the other hand, researchers like Auer (2013), Deppermann (2008) and Dorleijn & Nortier (2013) provide comprehensive sociolinguistic research on Kiezdeutsch, namely on its users and their functions in terms of self-positioning, language crossing, prestige, and usage in urban contexts. They define ways speakers use Kiezdeutsch to challenge hegemonic and often discriminatory ideologies. Conversely, they reveal how Kiezdeutsch can be used against speakers through comedic portrayals which serve to reinforce these ideologies. These findings are useful for understanding the ways speakers position themselves and others through language practices.

Moreover, their research reveals the challenges that sociolinguists face attempting to provide a "clean" description of phenomena like Kiezdeutsch. For one, they often find it difficult to keep pace with its highly dynamic usage and contexts. As a result, they dismiss areas such as grammatical characteristics as "violations" (Dorleijn & Nortier, 2013) or evidence of a "uncertain command of German morphology" (Auer, 2013, p. 36).† These views suggest a bias towards standard German language ideologies in their research. As well, data collected in street interviews may not be entirely representative of the usage of Kiezdeutsch. Preseau (2018) warns of an **observers' paradox**, whereby speakers are prone to code-switching or **hypercorrection** in the presence of adults or strangers, as demonstrated by the interviews conducted by Wiese (2012). In addition, researchers are rarely congruent in their judgements on Kiezdeutsch's definition. For example, Dorleijn & Nortier (2013) consider Kiezdeutsch to be **language play**, a manipulation of a dominant base form with features of a second one. For the same reasons, however, Deppermann (2008) calls it a dialectal variety of German.

A neglected area that warrants further investigation is the usage of Kiezdeutsch as speakers reach adulthood. Dorleijn and Nortier (2013) question its usefulness when users move into adult roles such as jobs or family responsibilities; Auer (2013) implies its usage denies youth educational success and professional careers; both Deppermann (2008) and Wiese (2012) claim Kiezdeutsch appears in informal speech when adults are not present. For these reasons, they consider Kiezdeutsch a youth variety. However, it is at this focal point where its definition becomes an academic debate of an "ephemeral style" versus "emerging dialect". In any case, future studies, especially concerning usage and age, can help sociolinguistic researchers better understand its usage across varying formalities, or **registers**, and social contexts such as work or education; and in general, lead to a more uniform analysis.

Conclusion

Language matters are complex. They encompass not only linguistic areas, but social ones too. In this paper, I examined Kiezdeutsch, a controversial variety of German that emerged only in recent decades. I started from a historical standpoint to understand where it originated, who its speakers are, and why they use it. As well, I briefly looked at some of its prominent linguistic features that distinguish it from Stan-

dard German. Its predominant users, German youth, are wielders of this variety, experimenting with the linguistic repertoires they have access to and ultimately, dispensing it for a variety of reasons: to signal identity; for comedic purposes; or to participate in the increasingly multicultural, multilingual contexts they live in. However, it is in the wider context of German society that Kiezdeutsch is viewed members of the Standard German speech community as problematic.

Auer (2013) writes the debate of style versus variety is, above all, a matter of perspective; Kiezdeutsch is defined not only by researchers, but also by its speakers and society. It is also important to understand the research produced so far on varieties like Kiezdeutsch is discussed around galvanizing themes such as immigration and "foreignization", themes which many perceive as threats to the standard language ideologies intrinsic to German society. As such, he and many other researchers recognise the lopsided debate on Kiezdeutsch—one that caricaturises "self-described 'ghetto-dwelling *Kanaken*' on the fringes of society" who refuse to integrate into German culture, and ignores a reality where speakers do successfully embed themselves into the mainstream society (p. 10).[6]†

In encountering similar narratives, Quist's (2008) research on the similar Danish *koebenhavnsk* multiethnolect seeks to reconcile the argument of style versus variety. On the one hand, she asserts that arguments for *style* reflect a stronger emphasis on sociolinguistic aspects such as self-positioning or crossing. On the other hand, arguments for *variety* (or, in this case, *Dialekt*) mostly focuses on structural and lexical forms from which a general systematic pattern emerges. Quist (2008), Auer (2013) and Wiese et al. (2014) agree these two perspectives can potentially complement each other, leading to more "fruitful" analyses for future studies; they could give sociolinguistic researchers and non-linguists not only a more well-rounded description of non-standard linguistic forms and functions, but also a deeper understanding of the ways language ideologies in society shape our perceptions of such phenomena.

6 See Haruna (2011), a *fluter* magazine article on youth speech that provides its reader an idea of the various experiences speakers have with varieties like Kiezdeutsch in context (note: in German)

References

Auer, P. (2013). Ethnische Marker im Deutschen zwischen Varietät und Stil. In A. Deppermann (Ed.), *Das Deutsch der Migranten* (pp. 9-40). Berlin: Walter de Gruyter.

Barbour, S. (2000). Germany, Austria, Switzerland, Luxembourg: the total coincidence of nations and speech communities. In C. Carmichael, & S. Barbour, *Language and nationalism in Europe* (pp. 151-167). Oxford: Oxford University Pres.

Barbour, S. (2005). Contemporary English influence on German – a perspective from linguistics. In G. Anderman, & M. Rogers, *In and out of English: For better? For worse?* (pp. 153-160). Clevedon: Multilingual Matters Ltd.

Bezemer, J., & Yandell, J. (2019, November 27). *Semiotic Resources*. Retrieved from Glossary of multimodal terms: https://multi modalityglossary.wordpress.com/semiotic-resources/

De Velde, J. R. (2017, December). Temporal adverbs in the kiezdeutsch left periphery: Combining late merge with deaccentuation for v3. *Studia Linguistica: A Journal of General Linguistics*, 301-336. Retrieved October 22, 2019, from https://libkey.io/libraries/513/ articles/59486761/full-text-file?utm_source=api_572

Deppermann, A. (2008). Playing with the voice of the other: Stylized Kanakspak in conversations among German adolescents. In P. Auer (Ed.), *Style and social identities: Alternative approaches to linguistic heterogeneity* (pp. 325-360). Berlin, Germany: De Gruyter, Inc. Retrieved September 14, 2019, from http://ebook central.proquest.com/lib/sfu-ebooks/detail.action?do cID=364724

Dorleijn, M., & Nortier, J. (2013). Bilingualism and youth language. *The Encyclopedia of Applied Linguistics*, 1-7. doi:10.1002/9781405198431.wbeal0103

Freywald, U., Mayr, K., Özçelik, T., & Wiese, H. (2011). Kiezdeutsch as a multiethnolect. In F. Kern, & M. Selting, *Ethnic styles of speaking in european metropolitan areas* (pp. 45-73). Amsterdam, The Netherlands: John Benjamins Publishing Company. Retrieved October 15, 2019, from http://ebookcen tral.proquest.com

Haruna, H. (2011, June 20). Weissu - is krasse Sprache! In der Clique, auf dem Schulhof, im Viertel: überall kann eine eigene Sprache entstehen. Ein Lauschangriff. *fluter*(39). Bundesz entrale für politische Bildung. Retrieved November 14, 2019, from https://www.fluter.de/weissu-is-krasse-sprache

Hilgendorf, S. K. (2007). English in Germany: contact, spread, and attitudes. *World Englishes, 26*(2), 131-148.

Hinnenkamp, V. (2005). Semilingualism, double monolingualism and blurred genres - on (not) speaking a legitimate language. *Journal of Social Science Education*, 57-90. doi:https://doi. org/10.4119/jsse-340

Preseau, L. D. (2018). Kiezdeutsch, Kiezenglish: English in German multilingual/-ethnic speech communities. (*Doctoral Disser tation, University of California, Berkeley*). Berkeley, California, USA. Retrieved September 6, 2019, from http://digitalassets.lib.berkeley.edu/etd/ucb/text/Preseau_berkeley_0028E_18133. pdf

Quist, P. (2008). Sociolinguistic approaches to multiethnolects: Lan guage variety and stylistic practice. *International Journal of Bilingualism*, 12, 43-61. Retrieved from https://libkey.io/li braries/513/articles/5428312/full-text-file?utm_source=api_572

Stevenson, P. (1997). *The German-speaking world: A practical intro duction to sociolinguistic issues*. New York, NY, USA: Routledge.

Stevenson, P. (2017). *Language and migration in a multilingual metropolis: Berlin lives*. Cham, Switzerland: Palgrave Macmil lan. doi:https://doi-org.proxy.lib.sfu.ca/10.1007/978-3-319- 40606-0

Wardhaugh, R., & Fuller, J. M. (2015). *An introduction to sociolinguistics* (7th ed.). West Sussex, England: Wiley-Blackwell, John Wiley and Sons.

Wiese, H. (2012). *Kiezdeutsch: Ein neuer Dialekt entsteht*. Munich, Germany: C. H. Beck.

Wiese, H., & Tanış Polat, N. (2016). Pejoration in contact. In R. Fink beiner, J. Meibauer, & H. Wiese, *Pejoration* (Vol. 228, pp. 243-267). Amsterdam, The Netherlands: John Benjamins Publishing Company. doi:10.1075/la.228

Wiese, H., Simon, H. J., Zappen-Thomson, M., & Shumann, K. (2014). Deutsch im mehrsprachigen Kontext: Beobachtungen zul exikalisch-grammatischen Entwicklungen im Namdeutschen und im Kiezdeutschen. *Zeitschrift für Dialektologie und Linguistik*, 274-307. Retrieved September 15, 2019, from https:// www.jstor.org/stable/43821602

Alight
Sayaka Hirano

Can Eliminative Materialism Account for Our Perceptions?

Costanza Saettoni
Università di Siena

In this paper, I aim to defend the theory of Eliminative Materialism as advanced by Paul and Patricia Churchland, against objections regarding our perceptions. I will do so by bringing forward cases of illusions. Based on anthropological and psychological studies concerning the Müller-Lyer illusion, it is entirely possible to maintain that our way of seeing is dependent on our cultural background. I will thus argue that Churchland's thesis is not as implausible as it may at first appear.

The theory of eliminative materialism (or *eliminativism*) was elaborated by Paul and Patricia Churchland in the early 1980s, in the context of the debate over the mind-body problem. It was proposed as a new monistic approach, situated on the complete opposite side of the dualistic view of Cartesian descendance, which views mind and body as two separate entities. The theory offered an alternative solution to the linguistic model first advanced by Chomsky and to Putnam's functionalism, based on the analogy between the working of the brain and that of the hardware of a computer. Although the theory remains still within the same category of materialism as the others, it is fundamentally linked to the theory of connectionism, as elaborated by the neuroscientists at the University of San Diego—in particular, the PDP Research Group. As with other forms of materialism, the first great advantage of this theory is that it does not require a commitment to sense-data[1] or other metaphysical objects: only what is physical exists in reality. According to the theory, the processes that happen in the

1 The term sense-data refer to mind-independent things that are immediately known in sensation; sensation, according to the theory, is when we become aware of such things. (Russell, 1912/2015, p. 11)

brain of a person have no one-to-one correspondent to the mental states we assign to the subject. This is not because of some sort of concession to dualism, but rather because we employ an inaccurate language that does not refer to concrete things, namely that of folk psychology, according to which we attribute specific propositional attitudes (e.g. belief, emotions, perceptions) to other people based on their external behaviour and gestures.

In his 1979 book, *Scientific Realism and the Plasticity of Mind*, Paul Churchland argues, among other things, for the theory ladenness of all empirical observations, which means that all our observations depend on our theoretical background. From this assumption, it follows that if we are able to change our theoretical framework—in particular, Churchland refers to that of folk psychology—we can change the phenomenology of our perceptions, not only of the external world, but also of introspection of our mental states. This belief in our capacity to extend our introspective ability, however, faces a serious objection. It has been argued that the claim lays its foundation on a confusion between perceptual judgments (i.e. how we describe the external world and our mental states) and perceptions (i.e. the way in which we come in contact with information about the world), and the idea that the two are based and dependent on the same theoretical framework. Our brain is the way it is as a result of biological evolution: it seems unlikely that we will ever be able to see sounds or hear colours (Nannini, 2011). Even though we are able to change our theoretical framework, it does not seem to necessarily follow that our perceptions of the external world would change with it, as Churchland claims.

In this paper, I aim to defend Churchland's theory by bringing forward cases of illusions. Based on anthropological and psychological studies concerning the Müller-Lyer illusion, it is entirely possible to maintain that our way of seeing is dependent on our cultural background. Therefore, I will argue that Churchland's thesis is not as implausible as it may at first appear.

Eliminative Materialism by the Churchlands

Paul Churchland entered the debate on the mind-body problem through his essay (1979), in which he brings forward an alternative theory to the mind-brain identity theory. The core of the theory remains entirely materialistic: all that exists is what is physical. The breaking point with the other materialist theories lies in the fact that

mental states—such as fear, love, but also propositional attitudes such as belief—as described in our ordinary language, cannot be reduced to cerebral states, i.e. the actual, physical states the brain is in. The impossibility of this reduction is due to the fact that our ordinary language is too imprecise: it is a legacy of a Cartesian, dualistic view of the mental. The gist of the argument is that folk psychology is wrong: and since it is just one possible theory, it can be discarded and replaced with a more precise one, just like theories in other scientific fields have been replaced through the centuries. Additionally, according to Churchland it is not only folk psychology that influences our perception, but also our theoretical and cultural background as a whole.

The reductionist approach was embraced by Churchland through Feyerabend; in particular, in his essay he cites the article "Explanation, Reduction, and Empiricism" (1962). In this article, Feyerabend actively argues that any materialistic theory would prove, in the end, that common sense psychology is wrong. To this view, Churchland adds the naturalized epistemology of Quine (1969), according to whom there are no theory-free objects of knowledge, and therefore it is impossible to reduce a scientific assertion to pure object-data. Another fundamental influence on the elaboration of the theory is Sellars' criticism of the 'myth of the given', explained at length in the paper "Empiricism and the Philosophy of Mind" (1956). The argument attacks the theories, such as sense-data theories, that claim that our knowledge is based on a set of "given". On this basis, Churchland (1979) claims that there are no empirical observations that do not presuppose a theoretical framework; even our 'introspective' perceptual judgements, as well as our perceptions, are bound to be theory laden (Nannini, 2011). In this regard, he writes in his essay (1979, p. 24):

> "The conviction that the world instantiates our ordinary observation predicates cannot be defended by a simple appeal to the "manifest deliverance of senses". Whether or not the world instantiates them is in the first instance a question of whether the theory which embeds them is true, and this question in turn is primarily a matter of the relative power and adequacy of the theory as a means of rendering the world intelligible."

As previously mentioned, folk psychology is a theory that elaborates the information we collect from our internal perception in line

with a certain image of human nature, resulting from a determinate historical tradition. Since folk psychology is a theory—and a wrong one at that—it is possible to change it; and with it, we can eliminate its ontological commitments. The idea of the necessity of a scientific revolution, capable of changing the way we see the world, is clearly indebted to Kuhn's epistemological theories, as put forward in the 1962 essay, *The Structure of Scientific Revolutions*, as well as to the crucial claim that no experimental data are independent from our theories.

Eliminative materialism heavily relies on the development of Artificial Neural Networks (ANN, or connectionist systems), processing systems of information inspired by the nervous system, originally based on the works of McCulloch and Pitts in 1943 (Graupe, 2013, pp. 9-10). In particular, the cornerstone for the Churchlands was the publication of the essays on the PDP (Parallel Distributed Processing) model in 1986 by D.E. Rumelhart, J. L. McClelland and the PDP Research Group in San Diego, where the two philosophers were working. This breakthrough was fundamental for the elaboration of an alternative to the functionalist mind-computer analogy. The new model tries to imitate the physiology of the brain: it consists in an apparatus of small elements, called units, that process inputs; each unit interacts with the others by sending exciting or inhibitory signals to others. This process aims to describe how information is processed by the brain, hence why the structure closely resembles the connections of the neurons in the brain. The passage of information through the brain as explained by the model plausibly accounts for how, even for ordinary tasks such as grabbing a cup, we constantly have to use a lot of information at once. The model is also able to give a plausible account of human thought, in that PDP models reflect the sequential quality of our cognition, i.e. the fact that when we form thoughts, we do so in a sequential series of transitions. (Rumelhart et al., 1986) The many advantages produced by the theory offer a good basis for eliminativism, as opposed to linguistic models and classic artificial intelligence, and it helped eliminativists to be taken more seriously. The neuroscientific theory thus provides sufficient grounds to dispute the model of classic artificial intelligence and Chomsky's suggestions. There are quite a few issues that symbolic theories of cognition (i.e. theories according to which the brain functions in the processing of symbols similar to a Turing machine or to a serial digital computer) cannot account for. One such problem is

that perceptions are bound to multiple sources of information. When a subject recognizes an object, many different cues are simultaneously processed by the brain; these stimuli could also be interpreted in different ways by the brain. When one tries to pick up the object, a lot of information is taken into account at the same time: shape and weight of the object, distance from the subject, location in the environment, and all these factors are necessary for the act of actually reaching the object with the hand. The production of language is constrained by sets of rules and the context in which it is produced: researchers in A.I. have thus proposed certain "scripts" to describe what we need to know in order to guide processing. Supporters of the PDP model, on the other hand, argue that interpretation of everyday events needs employ different kinds of knowledge. If we accept a modular processing system[2], it is difficult to account for the needed sensitivity to different sources of information; such models would have to simultaneously provide not just one, but several scripts in order to explain the interpretation of everyday events, which does not reflect the simplicity in which we ordinarily act.

Eliminative materialism thus offers a very plausible replacement to the language-based model and can find substantial support in neuroscientific theories.

Perception in Eliminative Materialism

The central feature of perception, according to Churchland, is that all empirical observations, both introspective and of the external world, are theory dependent. Since birth, we learn from others to perceive the world as everyone else does; consequently, Churchland argues, there is the possibility that we may learn to perceive the world in a way that is different from what our present culture teaches us to (1979, p. 7). Perception consists in how our brain uses the information it receives through the senses: what Churchland asks himself, then, is how good our exploitation of this data is. Scientific discoveries, he argues, have changed the way we perceive the world. The main example he offers in (1979, pp. 16–25) is that of caloric, which is essentially an elaboration of how people's perception of heat changed with progress in the field of thermodynamics. The common sense theory claimed that caloric was a substance that belonged to bodies, and which gave them

2 Idea that a given process operates only on its direct inputs and is unaffected by the operations of other modules. (Rogers et al., 2014)

their distinctive heat. The kinetic theory, however, has revealed that such entity does not actually exist: rather, heat is a form of motion. The argument is that the new theory has a much better explicatory potential compared to the theory of caloric: other than the fact that it has a better feedback from empirical data, it is also advantageous to substitute the archaic view with the new one. If we substitute folk psychology with neuroscience, then, our perception both of internal states and the world changes as well. Churchland also explains how one can change their perception of the night sky so as to actually *see* what Copernicus' theory implies—namely, that the stars we see in the night sky do not revolve around the Earth (1979, pp. 30-35)[3].

The Epistemological Aspect of the Theory

The form of eliminativism advocated by the Churchlands is fundamentally based on the epistemology of the post empiricists, according to whom there exists no completely theory-free observational language. More precisely, the cultural settings of the observer are fundamental for empirical observations: the way we see the world is determined by the way our brain organizes the flux of information it receives via our senses. Our knowledge depends entirely on our observations: since these necessarily presuppose a theoretical background, they are as fallible as theories themselves. But just as we can change theories, so can we better our exploitation of empirical data.

The PDP model offers a greatly improved theory for the explanation of how our brain works. The language-based model endorsed by Fodor—but which has its roots in the beginning of philosophy itself—is much too connected to an anthropocentric study of consciousness and brain. Given the fact that our brain is the result of a biological evolution that started with primates, it is hard to accept that our brain works in an entirely different way than that of other animals, by following linguistic rules to process external inputs: it seems far more plausible that the language centres of the brain are the result of a recent evolution, but that the brain essentially works under different rules (Nannini, 2011).

---

3 This is done, according to Churchland, by acquiring two elements , where the preliminary one is, of course, to know and understand the Copernican theory. The first element is then to learn to recognise by sight the planets of the solar system. The second element is to learn to conceive the movements and positions of the planets in a different—although still natural—way, that is, in a new coordinate visual system that takes as its ground line an elliptic plane (Ibi.).

Sellars (1956) offers a significant argument against foundationalism in epistemology. His criticism is directed against the "Myth of the Given", which refers to the notion that our knowledge is grounded on 'given' items: sense-data. The rejection of sense-data to account for our perception means that philosophers that followed in Sellars' footsteps do not need to appeal to metaphysical objects.

The Problems of the Phenomenological Aspect[4]

Churchland (1979) describes how we are able to change our view of a starry night so as to have the visual perception of the truth of Copernicus' theory. In the same way, our introspective perception allows us to know that we have certain desires and beliefs. That, however, is only an interpretation that we give to the information we gather according to a certain theory—namely, folk psychology. Since this is only one possible theory, if we substitute it with neuroscience, we should be able to change how we interpret our introspective sensorial data, and consequently also external data.

Objections to the Theory

There is a serious objection that the theory has to face in regard to the phenomenology of perception, in particular for what concerns our alleged ability to expand our introspective capacities once we get rid of folk psychology (Nannini, 2011, pp. 203–206). What the opponent may ask is: has the (almost) universal acceptance of the veracity of the heliocentric theory of the universe changed the way in which we perceive a starry night? Or do we not see the same night sky that Ptolemy saw? It is hard to accept, one can argue, that the way in which we come in contact with information about the world depends on the same theoretical-cultural background as the way we describe the external world and our mental states. In fact, Churchland seems to confuse three different theories (Nannini, 2011, p. 204):

- that our perceptual judgements depend on the language and psychological theory we use, and a change in the latter would also change our judgements;
- that our perceptions give us an image of the world dependent on the way our brain organizes the flux of information from our senses; it is therefore possible to imagine animals or aliens with different perceptions than ours;

4 In other words, the "phenomenological aspect" refers to "what it is like to perceive" (Fish, 2010, p.2).

- that our perceptions depend on the same theoretical framework as our perceptual judgements, and they change with the transformation of this framework.

While the first two hypothesis are more than plausible, the last is objectionable. It seems absurd to think that we could be able to perceive the world through echolocations like bats do, simply by changing our background scientific theory. Our brain is the way it is as result of millenniums of years of evolution: how can a new theory, even a more precise one, change our perceptions of the world—and of our inner states? Assuming, for example, we find a new theory that is able to describe in precise terms the electromagnetic waves that generate each colour, our perception of a red object would still remain the same, even though we would be able to describe it in different, more scientific terms. In short, the theory builds on a confusion between 'theoretical constructs' of psychology[5] and the 'mental constructs' of our brain, such as ideas and thoughts. Perceptions are elaborated based on a natural background, whereas perceptual judgements depend on a cultural background.

A further objection that can be advanced to the theory is that it is far too ambitious. It pretends, through a cultural and theoretical revolution, to change not only how we describe our own perceptions about the world and our internal states, but also the very way in which we perceive. According to the theory, abandoning folk psychology means that we can use more accurate, neurophysiological terms in order to describe our internal states: but still, is that enough to change how we perceive?

A Possible Reply to the Objection. The Müller-Lyer Illusion

The objection to Churchland's theory is very well founded. There are, however, a few possible counterarguments in support of eliminativism. I will focus on studies on the Müller-Lyer illusion that show how its strength depends on our cultural background: people with different backgrounds suffer a different susceptibility to it. Thus, perceptions

5 The term refers to the constructs that derive from the different theories of mind in psychology. For cognitive psychology, for example, the mind is like a computer that elaborates the information that arrive through sense organs and emits emotional and behavioral outputs, depending on the automatic assessment it does of the inputs; according to analytic psychology, on the other hand, outputs are the products of subconscious, of instinctual drives of the Id, of its inner conflicts and of the intervention of Ego and Super-ego. (Perdighe & Mancini, 2016).

can change with a change in our cultural and theoretical background.

The Müller-Lyer illusion has been used several times during the years by both Churchland and Jerry Fodor, in order to sustain their opposing claims. The main point of the debate between the two philosophers is whether we should accept a foundationalist account of justification in epistemology or not. The battlefield in which they argue is in regard to the question whether theoretical conception is capable of penetrating perception thoroughly. The two philosophers distinguish—albeit not very clearly—between diachronic penetration of perception, which lasts over a longer period of time, and synchronic, which instantly stops the susceptibility of a subject to the illusion (or, at least, lessens it).

Both philosophers accept the plasticity of mind and the theory-ladenness of some observations. Fodor claims, however, that it is possible to have theory-free observations, and the Müller-Lyer illusion is proof of the truth of the claim (Fodor, 1984). According to his interpretation, the persistence of the illusion is proof of the possibility of theory-free observation. His argument is that since the illusion is absolutely impossible to suppress, it follows that there are observations that are not theory-laden. On the opposite side, Churchland claims that "observational knowledge always and inevitably involves some theoretical presuppositions or prejudicial processing" (Churchland, 1988), and the Müller-Lyer illusion shows exactly this. Our theoretical background always has an influence on the way we see things. His hypothesis is that it is diachronic penetrability rather than diachronic encapsulation[6] which predominates perception, as it is also suggested by the studies on inverting lenses (see Kuhn, 1962).

Studies on the Müller-Lyer illusion

An interesting study analysed by McCauley and Henrich (2006) has brought a significant contribution to the debate. The study was conducted by anthropologists and psychologists in different communities around the world (Segall, Campbell, and Herskovits, 1966; based on studies by W. H. R. Rivers, 1901, 1905). The experiment consisted in submitting five illusions, among which the Müller-Lyer one, to people of different ages and with very different cultural and environmental

6 Encapsulation refers to the idea that information in the brain cannot be accessed by different psychological systems than the ones used to elaborate it. (Fodor, 1984)

backgrounds. The findings show that the Müller-Lyer is proof that perception is, in fact, diachronically penetrable.

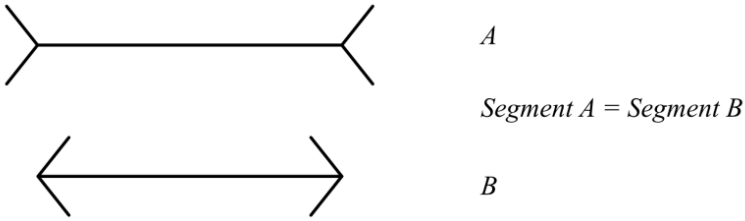


Figure 1 .

"Western" subjects are usually inclined to perceive the segment A (Figure 1) as longer than B. In the study, researchers varied multiple time the length of the two segments and each time asked the participants which of the two was longer. Then, they measured what length difference between the segments was necessary for the subjects to see them as equal. The strength of the illusion was defined by the estimation of this length difference: the longer the segment B had become, the stronger the susceptibility to the illusion. The groups interviewed are shown in Table 1 below.

Figure 2 displays the results of the study: on the left side, it is indicated the percentage of how longer segment A had to be than B in order for the interviewees to perceive them as equal (PSE score).

Group	Country/City	Sample Size
Ankole Adults/Kids	Uganda	131/93
Toro Adults/Kids	Uganda	49/37
Suku Adults/Kids	Congo Republic	40/21
Songe Adults/Kids	Congo Republic	45/44
Fang Adults/Kids	Gabon Republic	42/43
Bete Adults/Kids	Ivory Coast	38/37
Ijaw Adults/Kids	Nigeria	47/37
Zulu Adults/Kids	South Africa	21/14
San Adults	Kalahari Desert	36
S.A. European Adults	Johannesburg	36
S. A. Miners	South Africa	60
Senegal Adults/Kids	Senegal	74/51
Dahomey Kids	Guinea Coast	40
Hanunoo Adults/Kids	Philippines	37/12
Evanston Adults/Kids	U.S., Illinois	111/77
Bassari Adults/Kids	Eastern Senegal	50/50
Yuendumu	Central Australia	52

Table 1. Details for Samples (McCauley & Heinrich, 2006, 92)

On the right side, the vertical line shows the difference of the response of children and adults of the same group.

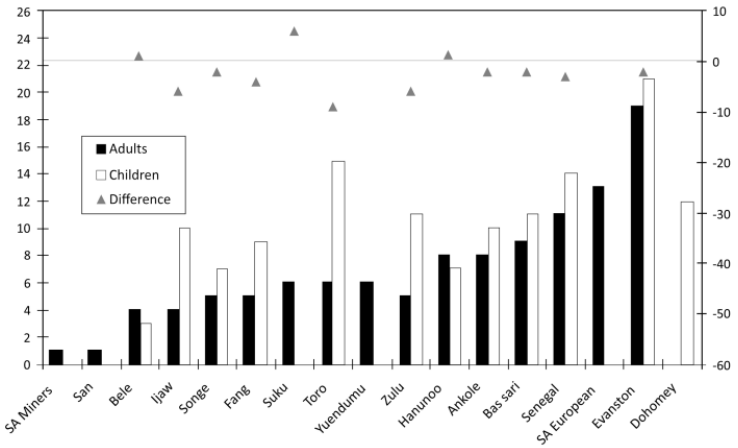


Figure 2. McCauley & Heinrich, 2006, 93

The sample of American adults in Evanston (USA) is the most susceptible to the illusion of the groups: segment B had to be approximately one fifth of the other for them to see the two lines as equal in length. Hunter gathers from the Kalahari Desert, on the other hand, are fundamentally immune to the illusion: A had to be just one percent longer than B for them to see them as having the same length.

The results among children, aged from 5 to 11, have a pattern similar to the one of the adults. The graph (Figure 2) shows that the differences between cultures is more pronounced between children than it is between the groups of adults. In the case of the children belonging to the Suku group, for example, ethnographers reported that PSE was 0%, whereas American children in Illinois scored a PSE of almost 22%, higher than the one of the adults of the same cultural group.

Moreover, further studies among American subjects (see Wapner & Werner, 1957) have proved that the susceptibility to the Müller-Lyer illusion decreases between the age of 5 and 12, and then increases again from 12 to 20, after which it does not change.

Segall, Campbell, and Herskovits' (1966) explanation for the difference in susceptibility to the illusion centres on the "carpentered environments" hypothesis: the difference is due to the prevalence in one's cultural environment of rectangular shapes, "a factor which seems to be related to the tendency to interpret acute and obtuse

angles on a two-dimensional surface as representative of rectangular objects in three-dimensional space" (1966). Thus, people who live in an urban environment such as Evanston in Illinois are more likely to fall for the illusion.

Conclusions

From this study we can draw three conclusions:

1. what influences people's susceptibility to the illusion works between birth and age 20, therefore susceptibility to the illusion is not innate;
2. the cause—or causes—have their effects mostly before age 11, or children's pattern in Figure 2 would not be similar to the adult's;
3. the focus should turn on what kind of experience before the age of 20 is the cause of the susceptibility.

The point of interest for the present discussion is that the study shows that the effectiveness of the illusion is dependent on the cultural background of the person: the illusion is diachronically penetrable, at least during a certain age. People who grow up in different cultures and environments have different perceptions. It can therefore be argued that, even illusions that seem impossible to avoid, can actually be overcome, at least until a certain age: by changing the environment, our perceptions can change. This shows that our visual perceptions are never completely objective, and that our cultural background defines the way in which we perceive the world.

I believe that it is possible to use these results in order to answer the objection put forward before. The dependency of our perceptions to the environments means that, with a change in the latter, the former undergoes a change, too: it does no longer seem so implausible that a modification of our theories affects the way we perceive the world and our internal states. Once we accept that even illusions such as the Müller-Lyer—which Fodor frequently used as an example of an impenetrable illusion—can actually be suppressed, it becomes much easier to believe in the possibility that perception can actually change with a shift in our theoretical and cultural framework.

Discussions

I believe that the Müller-Lyer illusion offers a significant support to Churchland's claims. Studies on how the illusion works are good evi-

dence that susceptibility to the illusion is in fact related to the cultural and environmental background of a person. Although they may show that the plasticity of mind is not as strong as Churchland claims it is, it still remains an important proof in favour of the theory that empirical observations are theory dense. If it is possible to have different responses to this illusion, it could be that we might be able to change our perceptions through a modification of our background theories. As it is shown in the study, a different background implies different susceptibility to the perception of the world. If you change the former, it seems that also the latter is very likely to change. It is actually possible, then, that our perceptions are not independent from our background.

I believe that eliminative materialism offers a significant explanation for the mind-body problem, and it should not be overlooked too quickly. The advantages of accepting this theory are manifold. First of all, neuroscience appears to give substantial support to the claims of the theory. Furthermore, eliminativism description of the inner workings of the brain offers a more suggestive explanation than the language-based model endorsed by Fodor. With the PDP model, we do not need to resort to syntactical rules to account for cognition. Moreover, it is undeniable that folk psychology is a theory that has not evolved in centuries. It is not implausible that it might be wrong: by recognizing that it is not impossible to get rid of it, we can substitute it with a much better theory, which has a greater explanatory potential. Moreover, the radically materialistic view of the theory allows us to abandon the Cartesian dualistic ideal that has shaped the debate on the connection between mind and brain.

References

Churchland, P. M. (1979). *Scientific Realism and the Plasticity of Mind*. Cambridge, Mass.: Cambridge Univ. Pr.

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes. *The Journal of Philosophy*, Vol. 78, No. 2, pp. 67-90.

Churchland, P. M. (1985). Reduction, Qualia, and the Direct Introspection of Brain States. *The Journal of Philosophy*, Vol. 82, No. 1, pp. 8-28.

Churchland, P.M. (1988). Perceptual Plasticity and Theoretical Neutrality: A Reply to Jerry Fodor. *Philosophy of Science*, Vol. 55, No. 2, pp. 167-187.

Feyerabend, P. K. (1962). Explanation, Reduction and Empiricism. In: H. Feigl & G. Maxwell (Ed.), *Crítica: Revista Hispanoamericana de Filosofía*, pp. 103-106.

Fish, W. (2010). *Philosophy of Perception. A Contemporary Introduction.* New York: Routledge.

Fodor, J. A. (1984). Observation Reconsidered. *Philosophy of Science*, Vol. 51, No. 1, pp. 23-43.

Fodor, J. A. (1988). A Reply to Churchland's "Perceptual Plasticity and Theoretical Neutrality. *Philosophy of Science,* Vol. 55, No. 2, pp. 188-198.

Graupe, D. (2013). *Principles of artificial neural networks.* Singapore: World Scientific Publishing Co. Pte. Ltd.

Kuhn, T. S. (1996). *The Structure of Scientific Revolutions*. Chicago, Mass.: University of Chicago Press. (Original work published 1962).

McCauley, R. N., & Henrich, J. (2006). Susceptibility to the Müller-Lyer Illusion, Theory-Neutral Observation, and the Diachronic Penetrability of the Visual Input System. *Philosophical Psychology,* 19:1, pp. 79-101.

Nannini, S. (2011). *L'anima e il corpo. Un'introduzione storica alla filosofia della mente*. Bari: Gius. Laterza & Figli Spa.

Perdighe, C., & Mancini, F. (2016). *Elementi di psicoterapia cognitiva.* Giovanni Fioriti Ed.

Quine, W. V. (1969). *Ontological Relativity and Other Essays.* New York: Columbia Univ. Press

Rogers, T. T., & McClelland, J. L. (2014). Parallel Distributed Processing at 25: Further Explorations in the Microstructure of Cognition. *Cognitive Science*, 38: pp. 1024-1077.

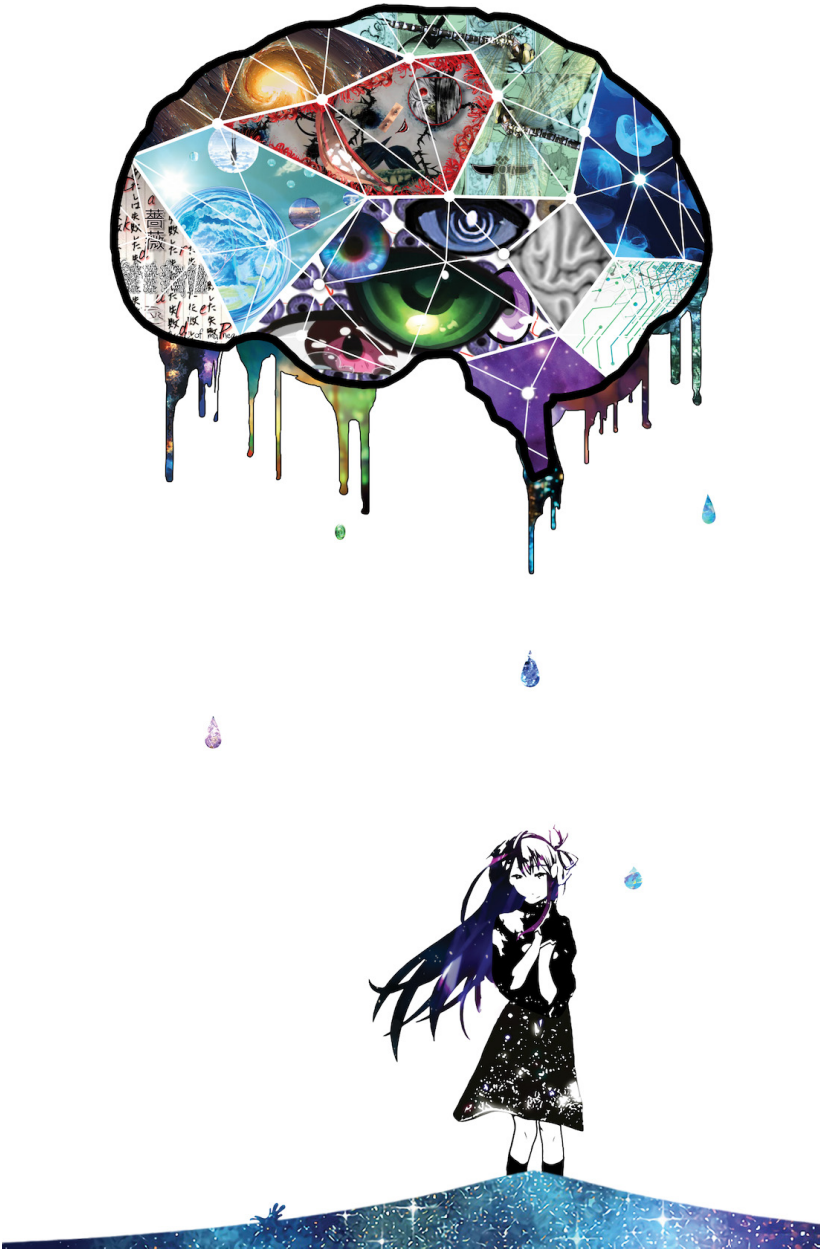
Rumelhart, D. E., McClelland, J. L. et al. (1966). *Parallel distributed processing: explorations in the microstructure of cognition. Volume 1. Foundations.* Cambridge, Mass.: MIT Press.

Russell, B. (2015). *The Problems of Philosophy*. New Jersey: J. P. Piper Books. (Original work published 1912).

Segall, M. H., Campbell, D.T., & Herskovits, M. J. (1966). Cultural Differences in the Perception of Geometric Illusions. *Science*, 139 (3556), pp. 769-771.

Sellars, W. S. (1956) Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of Science*, 1: pp. 253-329.

Wapner, S., & Werner, H. (1957). *Perceptual development: An investigation within the framework of sensory-tonic field theory*. Oxford, England: Clark Univer. Press.



Emergence ex Astra
Anonymous

Neuromodularity: A Potential Cross-Modular Consciousness

Larissa Melville
Simon Fraser University

We often ponder the nature of consciousness. Less common is to consider whether individual parts of conscious experience *combine* in the same way for every person. Are you able to recognize faces? Can you understand the words on this page? In recent years, some neurological conditions (e.g. prosopagnosia; this article will explore many) have illustrated that the stream of consciousness can be interrupted in meaningful ways—as if a piece of the conscious experience has been removed. Michael Gazzaniga (1988, 2018) proposes a modular sort of consciousness to explain these observations by pulling evidence from neuroscience research. His stance is thus coined here as *neuromodularity*. This article explains the facets of a potential neuromodular brain architecture: the history, the neurological construction of a module, and neuromodular interactions in a layered and parallel subsumption system. It explores how modular amplitude and presence are determined by the functionality of key brain structures, also explaining that by assuming neuromodularity, the conscious system does not disintegrate entirely at the loss of one module. Finally, this article pits additional neurological deficits against Gazzaniga's theory; the cross-sensory integration of vision and multimodal interventions for brain injury serve to test the integrity of a neuromodular system. It is theorized that despite their complexity, this framework supports them. Foremost, I intend to convince readers that Gazzaniga's theory of consciousness is worth pursuing as a valuable entry into the tome of consciousness literature.

Can someone ever be wrong about their conscious experiences? It is certainly possible. In fact, we have two blind spots in our vision because there are no photoreceptors where the optic nerve passes through the retina. However, this spot is typically unnoticeable in consciousness because the brain fills the gap with its best guess (Durgin, Tripathy, & Levi, 1995). This example suggests that other doubts about consciousness are possible, for instance, in the subjectivity of colour perception—a skill many believe is personally infallible. Surely, one cannot be wrong about their experience of red. But I think our confidence in this belief is unfounded; consciousness does not become more reliable over time and exposure. Rather, it is more like sleep: no matter how many hours are dedicated to the practice, we never get any better at it. In fact, I hope to convey in this article that the stream of consciousness is a ubiquitous deception.

This paper aligns with illusionism—a theoretical approach to consciousness which denies that experiences have subjective properties (Frankish, 2016), which are often called phenomenal properties of consciousness (i.e. the "what-it-is-likeness"; Nagel, 1974). One example is my experience of aged cheese; it has introspective and qualitative features: sharp, strong, and steady. Some illusionist researchers believe these phenomenal feelings arise from distortions of cognitive modules (see Marinsek, 2016). However, the issue of phenomenality is not the focus of this paper. Instead, I will offer an illusionist perspective which proposes that the constituent parts of phenomenal (should they exist) and cognitive systems produce individual pieces of consciousness; I will argue that the *unification of this modular activity* is an illusion.

This article will present a neuromodular take on consciousness pioneered by the cognitive neuroscience researcher Michael Gazzaniga (2018) in his book *The Consciousness Instinct*. However, I only scratch the surface of Gazzaniga's lifetime of research here. In particular, this article summarizes the proposal that consciousness is the result of parallel modular processing and some considerations involved with this claim. Gazzaniga likens the role of these modules to bubbles which project into consciousness their best explanation and interpretation of the world. He believes that modular activity slots together in a seamless illusion. Like a pot of boiling water, each bursting bubble is a piece of awareness: independent, complex, and stitched together by time. To explain how, this paper will investigate the brain as if it

has a layered architecture. Insights into brain damage and behaviour will bolster Gazzaniga's view and simultaneously discredit any global neural processor. The last section will address that I side with Gazzaniga because his framework recognizes data from cross-modal research and explains why split-brain patients sometimes cannot recognize the actions of their left hand. But first, it is imperative to consider a brief history of modularity.

Some Misguided Accounts of Modularity

For some, understanding the human experience is an onerous challenge; for others, it is the hardest problem of them all. Dan Dennett (2003, 0:55), a cognitive scientist and consciousness researcher, explains that when philosophers ask what he works on, "their lips curl into a snarl, and [he] get[s] hoots of derision and cackles and growls because they think: that's *impossible*." Like Dennett, I think they are wrong. Scientific evidence may yet explain the complexity of consciousness. Besides, the shadow it casts does not deter every philosopher, as many prominent views pepper the field, undeterred. In particular, the subject of modularity has grown over the past few decades, rising and falling in an adapting theoretical landscape. Likewise, Gazzaniga's (2018) addition to consciousness literature deserves some scrutiny. Although, I will first explore previous iterations of modularity to illustrate that Gazzaniga's perspective is compelling on many fronts.

To begin, the study of phrenology is one of the first known cases of physiological modularity, although not called so at the time. It was incepted in the early 19th century by Franz Joseph Gall. Phrenology is defined as the study of measuring skull protrusions to predict individual characteristics. Essentially, Gall believed that faculties of the mind resided in the cerebral cortex and were organized into 35 or so regions based on functionality. For instance, there were mapped areas for colour and form perception, and more fugacious qualities like cautiousness and secretiveness. When a region worked harder in comparison to others it "grew" and caused a bump on the cranium; it was thus a prominent quality of an individual's disposition. As an example, if someone was secretive, the region superior to the ear would be enlarged (see Figure 1; Gall & Spurzheim, 1810–1819; Gazzaniga, Ivry, & Mangun, 2014). The practice quickly got scathing reviews from Pierre Flourens (1846), as he found more accurate neural correlates for function (e.g. decorticate pigeons had no perception or motor ability).

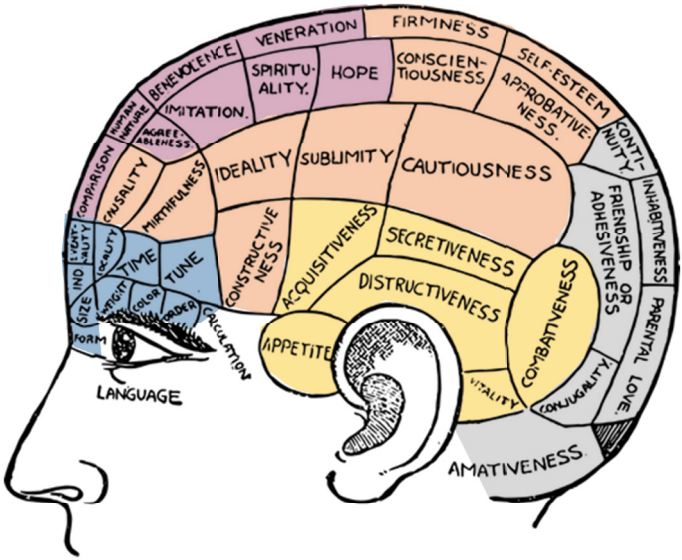


Figure 1. Definition of each phrenological organ

Note. This diagram locates the various phrenological organs on the scalp, such as secretiveness and mirthfulness.

If nothing else, phrenology seeded the idea of localization—that distinct modules in the brain could house individual functions.

The concept of modularity became popular at the dawn of Jerry Fodor's (1983) *Modularity of Mind*. Put simply, his theory suggests any cognitive system which largely falls into a set of nine constraints can be considered modular. For this article, the most relevant constraint is *information encapsulation*, which for Fodor, is weighted heavily above the others. Modules are restricted by the flow of information, both entering and exiting the module. Furthermore, Fodor believed that input systems such as perception were modular but central systems like reasoning were not. While Fodor's theory is widely disputed, especially on the point of information encapsulation (see Prinz, 2006), modularity has evolved as a useful tool to speculate about neurological and mental architecture.

Modularity, Again

The kind of modularity proposed by Gazzaniga (1988, 2018) is seated in neuroscience research. Gazzaniga defines a module as a specialized network of neurons that serves a particular purpose; there

may be hundreds if not thousands of modules available. It differs from the modularity proposed by Fodor (1983), which says that modules are components that contribute to certain cognitive functions, such as language (Gazzaniga, 1988). In Gazzaniga's (2018) framework, but not Fodor's (1983), language production is a possible module. Traditionally, language production was thought to be localized to the left inferior frontal gyrus in the pars triangularis and opercularis—a region called Broca's area (see Figure 2). However, it is almost never the case that a cognitive function can be isolated to one cerebral location. In fact, language production occurs beyond the scope of Broca's area; it is a complex network of white matter tracts which recruit parts of the premotor cortex and superior temporal gyrus, amongst others (see Figure 2; Friederici, 2015). Additionally, the primary auditory cortex is necessary to hear one's own speech and modify it appropriately (Hickok & Poeppel, 2007). The point is that Gazzaniga's modules are not spatially contiguous like a puzzle, in fact, many modules have multiple pathways, and may also share similar pathways. Modules should instead be analyzed though functional categories (e.g. language production).

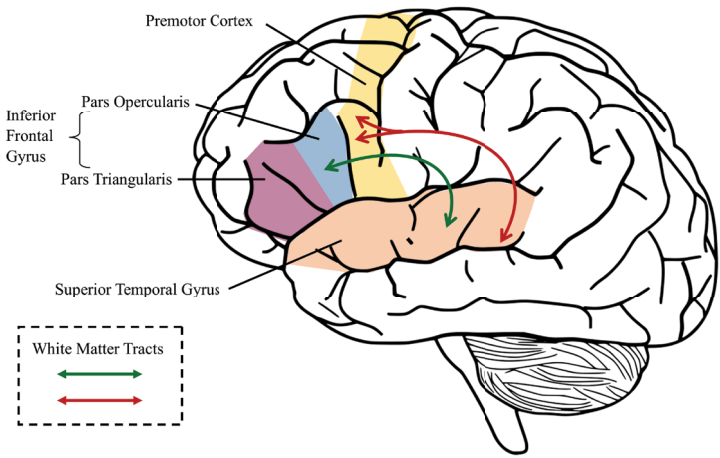


Figure 2. White matter pathways between the frontal and temporal language centres

Note. The red and green bidirectional arrows outline the two dorsal white matter pathways believed to activate during language production and preparation. The inferior frontal gyrus and Broca's area are sometimes synonymous; damage here may cause Broca's aphasia.

Crucially, this neuromodular architecture rejects the need for a global processor—a system where every neuron is connected. That is, there is no centralized system plucking consciousness from neural activity anywhere (Gazzaniga, 2018).

Gazzaniga (2018) explains that there are several advantages to a neuromodular set-up in comparison to a global processor. In one case, the energy cost of running a global system is impractical as dendritic connections would be too dense and axonal connections would be too long (Gazzaniga, 2018). Not only would this architecture slow processing speeds considerably, the brain would be twenty kilometers in diameter (Gazzaniga, Ivry, & Mangun, 2014; Nelson & Bower, 1990)! In a neuromodular system, a task can selectively activate relevant modules rather than the entire structure. In this case, axons are thinner and transmit electrical impulses faster, which helps complete tasks more efficiently (Gazzaniga, 2018). Another advantage to modular architecture is that concurrent tasks can be completed by different modules working independently. Simultaneous walking and talking serves as an example: the walking module (or multiple modules) is strained by additional motor tasks. It turns out completing a cognitive task is easier while walking than another motor task. For instance, discriminating between different tones would be easier to accomplish while walking than holding two sticks and not letting them touch (Beurskens, Steinberg, Antoniewicz, Wolff, & Granacher, 2016). In comparison, a global processor would be heavily overworked by doing these two simple tasks (Gazzaniga, 2018). Therefore, due to cost inefficiencies, slow processing, and the inability to process two simultaneous activities, a wholly neuromodular system is the more fitting architecture.

Finally, if neural processing was global, the entire system—consciousness included—would collapse at the loss of a module (Gazzaniga, 2018). Imagine any person who has suffered a stroke or brain lesion; they may still be able to see, hear, and walk. Sometimes after a stroke, the language production areas are damaged, which typically leads to a condition called *non-fluent* or *Broca's aphasia*. Non-fluent aphasia is an impairment of language production only; its absence does not beget total system failure. On the contrary, patients mostly retain cognitive function, phonemic articulation, and language comprehension (Conklyn, Novak, Boissy, Bethoux, & Chemali, 2012). Thus, the stream of consciousness continues, despite the lost module. This autobiograph-

ical consciousness is inherited only from the functioning modules, which may impede lost or deficient modules from entering conscious awareness.

More conclusive evidence for a non-global processor is found in studies of spatial hemi-neglect, an attentional condition caused by brain damage in which one half of a patient's world appears not to exist. Note this is not a deficit of vision, but rather of damaged attentional mechanisms typically in the right hemisphere (Behrmann, Watt, Black, & Barton, 1997). For instance, a patient might forget to put makeup on one side of her face, leave half of her plate unfinished, or fail to report details from parts of sensory stimuli. Furthermore, she will deny anything is wrong—a condition called *anosognosia* (Kerkhoff, 2001). In a test of extinction, where two objects are placed in a patient's two hemifields, only one of them is consciously registered (i.e. the ipsilesional side). However, if a singular object is presented, in either hemifield, the patient will readily recognize it (Gazzaniga, Ivry, & Mangun, 2014; Vallar, 1998). Gazzaniga (2018) believes this phenomenon describes a sort of competitive processing, where the functioning modules' activity overcomes any challengers. The losing module never makes it to consciousness. To reiterate, according to Gazzaniga (2018), system collapse would befall a global processor upon damage to it. By instead assuming neuromodularity, the piece that a module contributed to consciousness is either lost or diminished in amplitude when damaged, which supports the evidence that brain lesions do not shut down conscious awareness entirely.

Layered Architecture as a Neurological Avenue

Modules are not informationally encapsulated like Fodor (1983) thought. In fact, Gazzaniga (1988, 2018) emphasizes they are highly interconnected in the brain. Evidence comes again from studies of language neurology. To explain, another likely language module is semantic processing. Activation along a white matter tract from the superior temporal gyrus to parts of the inferior frontal lobe is associated with semantic processing (see Figure 2; Friederici, 2015). These fibers are a possible connection between two distinct and functionally similar modules. Modules that activate during a specialized process (e.g. language) tend to have stronger interconnectivity and interact less with unrelated modules (Gazzaniga, 2018). But how are these interactions structured holistically?

Gazzaniga (2018) recommends exploring the concept of a *layered architecture*, a term likely familiar to computer scientists and engineers alike. He suggests a similar architecture may occur in biological systems. To simplify greatly, layered architecture denotes a parallel processing system in which activation occurs simultaneously at all layers, rather than sequentially. While every layer communicates bidirectionally, processing may not omit a layer (e.g. layer 3 may output information to layer 2 but not layer 1), meaning information processed at lower layers is inaccessible to higher ones (Doyle & Csete, 2011; Gazzaniga, 2018). In a biological example, consider how exhausting it would be to be aware of every neuronal firing; producing language would incite thoughts for every phoneme, morpheme, syntactic rule, and so forth (if such a layered system occurred). To reduce energy consumption, the phonemes that are deciphered by the morpheme layer's protocol are hidden to the syntax layer. Essentially, the syntactic structure need not "know" of the phonemes.

Information is abstracted at each layer and the level of necessary detail depends on the processing required. Layers work on a "need to know" basis, behaving independently within a specific set of *protocols*—instructions that stipulate permitted communication at both an interlayer and intralayer capacity (Doyle & Csete, 2011; Gazzaniga, 2018). Layers are constrained like an hourglass; protocols are enacted at the neck, and granules of modulated outputs spread to the subsequent layers. Doyle and Csete (2011) find a layered system in clothing as an example. The primary function of clothing is protection from outside elements, but each layer must follow specific constraints. For instance, the t-shirt must be soft, the cardigan must provide warmth, and the jacket must deflect the rain. This construction allows for a limitless number of inputs to maximize the number of outputs (i.e. thousands of outfit permutations; Gazzaniga, 2018). There is an eerie similarity here between layered and neural architecture, as there are often many neural pathways that a single behaviour can activate (Gazzaniga, 2018). Perhaps, then, a layered architecture might delineate the complex processing between the modules of the brain.

If the brain is indeed layered, biological learning must be viably explained within that framework, although it is important to first understand learning in a layered architecture as it was originally proposed in robotics. That is, modules and layers overtake any need for a

centralized system. Energy costs notwithstanding, a global system like those in older robots would be prone to freezing at the presence of a previously unencountered stimulus (Gazzaniga, 2018). Rodney Brooks (1986) solved this problem by introducing *subsumption architecture*. Here, a layer is described as housing one or more behavioural modules. Simply put, novel information accretes new layers into pre-existing ones and becomes absorbed (subsumed) by the entire system. Individual nodes may run independently of each other, working on multiple goals concurrently, although, control can also be captured by higher layers, which subsume the roles of the lower layers (see Figure 3; see also Brooks, 1986, for more detail).

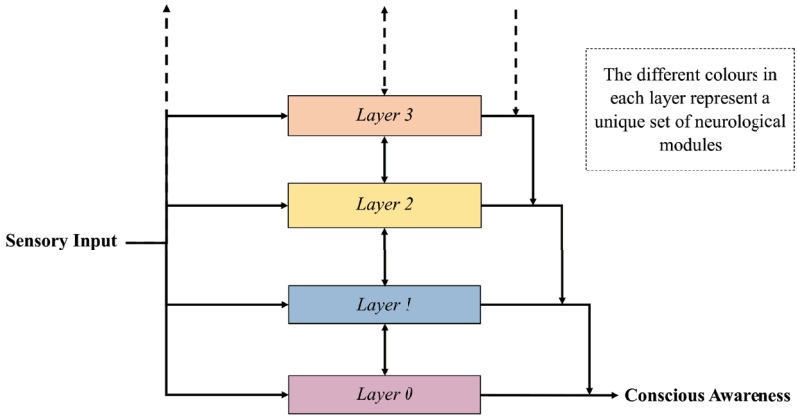


Figure 3. The process of subsumption in a layered architecture

Note. This is a potential representation of a layer network which projects its processes into consciousness. Each layer may work on individual goals concurrently. It is assumed here that communication is bidirectional between layers.

Although layered architecture is crucial for robot function, is it analogous in the brain? Gazzaniga (2018) believes it is. Recent biological research agrees with him. For example, there is speculation of a layered architecture in molecular communication (Nakano, Suda, Okaie, Moore, & Vasilakos, 2014). Yet, while the brain may be layered, the nuance and organization of layers remain a mystery. Gazzaniga (2018) proposes one potential layered learning system: a control layer which plans for future perturbations by extracting from memory and adjusting layer protocols as necessary. One version of this is sensitization, a learning which occurs when a repeated stimulus elicits stronger

responses over time (Overmier, 2002). I imagine getting bitten by a snake might cause a greater avoid response the next time one is encountered. In this case, the protocol has shifted from "explore the environment" to "explore the environment and watch out for snakes." In the end, Gazzaniga (2018) concedes that neuroscientists are nowhere close to solving how the brain regulates its layers. However, his stance may guide future neuroscience research.

Bubble, Bubble

There has been much preamble so far about modules, layers, and brain lesions, but how does it all construct consciousness? Gazzaniga (2018) hypothesizes the following: consciousness is derived from the activation of independent modules amalgamated into a comprehensive system, which is modulated only by the relative dominance and amplitude of module activity. As explained previously, modules are organized into complex layers to perform specialized functions, but each module is largely unaware of other modules' processing. By operating in parallel, neuromodular activity can appear simultaneously. Consciousness is a stream of activations from relevant modules, sewn together seamlessly by the brain in temporal unison. The fact that we are unaware of neuromodular fusion at a conscious level is the illusion of which Gazzaniga (2018) speaks. Lose one module, and that piece of consciousness is lost too, although, some modules might function under specific conditions (e.g. spatial hemi-neglect). Chiefly, it is almost never the case the consciousness is lost completely.

Evidence for the above hypothesis arrives from split-brain patients who appear to possess two independent conscious systems. To explain, split-brain patients have a severed portion of the corpus callosum, the large band of axons which connects the two hemispheres. It is also possible to have a commissurotomy, which is a complete disconnection of the corpus callosum plus all the remaining white matter tracts between hemispheres (commissures). In such cases, researchers have noticed that not only was the brain split, but so was the mind (Gazzaniga, 2018; Sperry, 1969). Since language is predominantly a left hemisphere function, when the brain is split, the right hemisphere has no capacity to speak. Early research on corpus callosotomy and commissurotomy patients show similar abnormalities in response to input in both visual fields. Either type of patient can freely verbalize items presented in the right visual field (perceived by the left hemisphere).

However, they fail to identify stimuli presented to the left visual field (perceived by the right hemisphere) because information transfer from the right hemisphere to the language module in the left hemisphere is gone. In fact, these patients deny seeing anything in the left visual field. In comparison, when asked to point with their left hand (controlled by the right hemisphere), they can pick an item matching what was presented to the right hemisphere (Gazzaniga, 2018; Sperry, 1970). It is as if each hemisphere has no conscious awareness of the other.

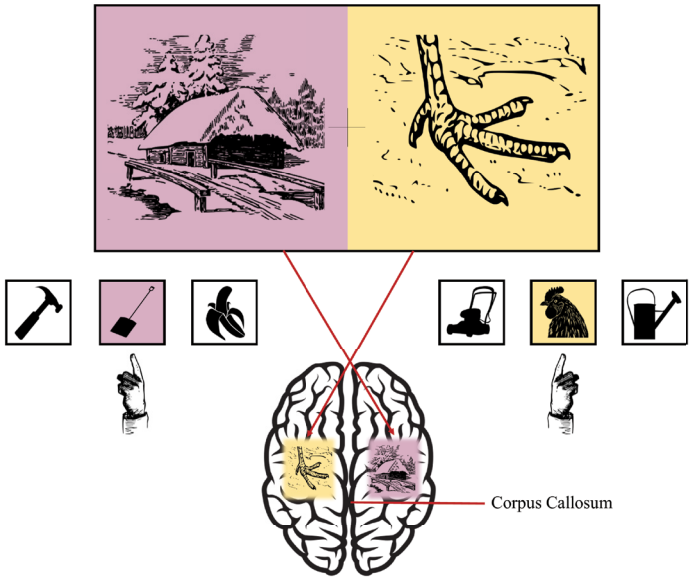


Figure 4. Neural processing of visual input in a split-brain patient

Note. A split-brain patient sees two different images in each hemisphere. The left hemisphere processes the chicken foot and points to a chicken with the right hand. The right hemisphere processes a snowy scene and points to the shovel with the left hand. The left hemisphere interpreter attempts to explain the shovel into the chicken narrative because it does not have access to the snowy scene.

A modular make-up might explain the perceptual irregularity seen in split brain patients. The left hemisphere is a major "interpreter" of its perceptions, that is, it attempts to find causal relations and explanations from its isolated information (Gazzaniga, 2000, 2018). In Gazzaniga's (2000) experiment, a split-brain patient is presented two dissimilar visual stimuli, one to each hemisphere, and asked to point to a related image for each stimulus from a randomized array. For example, the patient sees a snowy house in the left hemifield and a chicken

foot in the right. The left hand chooses an item based on what the right hemisphere sees—such as a shovel, which goes with snow. However, the left hemisphere does not know why the left hand is pointing at a shovel when it saw the chicken foot (see Figure 4). Intriguingly, the left hemisphere interpreter attempts an explanation: "oh, you use the shovel to clean out the chicken coop," fabricating a narrative congruent with the chicken scene. Gazzaniga (2018) explains that in split-brain patients, the left hemisphere contains modules that are inaccessible to the right, and vice versa. At some level, the stream of consciousness bifurcates, and two separate minds are born.

Gazzaniga (2018) likens his version of consciousness to a boiling pot of water. Each module has the capacity to appear in consciousness, and when activated, a little bubble will sprout to the surface. Upon breaking, burst of consciousness emerges, only replaced by another module's bubble in a timeless dynamic motion. The transition from one bubble of consciousness to the next is seamless—an illusion unbeknownst to most.

Braille, Music, and Magic

I want to present three accounts of neurological data which settle comfortably within Gazzaniga's (2018) framework. The first two concern the possible neuromodular and layered architecture of the brain, beginning with how modules have a capacity for cross-sensory integration, and followed by how a multi-modal therapy may outline a potential layer that travels across hemispheres. Finally, in the grand scheme, I will look deeper into the split-brain patient and offer an example of how neuromodular unity is illusory.

I must stress that Gazzaniga's (2018) definition for modules is deliberately flexible. It leaves the door open for phenomena like complex interactions, neuromodular overlap, multiple realizability, and cross-modal bubbles of consciousness; it allows for the mental resilience that slews of brain lesion research profess to. In one study (Merabet et al., 2008), dormant neural pathways in the occipital lobe were activated by touch after a period of vision loss. Participants were all sighted. In the experimental group, participants were blindfolded for a period of six days and periodically tested with fMRI. During this time, the participants learned braille and played other tactile games. Crucially, in comparison to the previous fMRI results, the fifth day showed strong activation in the occipital lobe. These data suggest that mod-

ule(s) in the visual cortex have pre-existing cross-modal connections that selectively activate for non-visual tactile stimulation. Gazzaniga's (2018) framework accounts for this data: if one pathway is lost, another will pick up the slack. Every pathway would need to be broken to lose "vision" (or consciousness) entirely. Notice this example characterizes *multiple realizability*—layered biological architectures often have many ways of accomplishing the same task. This suggests that the brain can generate sight in more than one way, be it through visual pathways or tactile ones, which supports a highly adaptable and robust system.

It remains unclear in Merabet et al.'s (2008) study whether a new module has been formed, or if an old module has been recruited. Therefore, I propose that a greater severity of module loss, whether lesioned or through disuse, will recruit functioning neural connections into a new module. Evidence comes from homotopic connections found between Broca's area homologs in each hemisphere, which I believe are two strongly connected modules in a layer. By extension, this neural connection serves to improve language production after Broca's area is lesioned. To explain, since 1973, speech-language pathologists have used intact musical ability as a tool for patients with non-fluent aphasia to recover language function. The technique was coined Melodic Intonation Therapy (MIT; Albert, Sparks, & Helm, 1973). These melodic features of language are found predominantly in the right hemisphere (Jeffries, Fritz, & Braun, 2003), and may be unaffected after a left hemisphere stroke. For instance, many patients with non-fluent aphasia can still sing the words to popular childhood songs unimpeded (Yamadori, Osumi, Masuhara, & Okubo, 1977). The effect of recruiting right hemispheric structures to support language is well-documented. In fact, participant scores on language tests can improve after just one session of MIT (Conklyn et al., 2012). Yet, it remains unclear what neurological change MIT incites. Either Broca's area is moderately restored to its original state or the right hemisphere takes control of language; however, current research suggests both hypotheses are correct. The type of plasticity depends on lesion severity. The right hemisphere only takes over when lesion severity is absolute (Schlaug et al., 2010), that is, a new module is formed. In the case where perilesional tissue around Broca's area can be repaired, the connection to its homolog is an example of multimodal fortification, where the old module regains some function. I believe the addition

of new modules or the recuperation of old ones is explainable within Gazzaniga's (2018) framework.

Gazzaniga (2018) claims that modular activations each have the capacity to make us aware of their processing. They are linked together through time like a movie shown in 24 frames per second; the transition between frames is invisible to the eye much like the transition between the parallel neuromodular processing that makes up consciousness. This neuromodular unity is not damaged in split-brain patients. When the two hemispheres are in conflict, the left hemisphere is confused but not isolated by the right hemisphere's behaviours (Marinsek, 2016). Post surgery, some patients make comments about their left hand (controlled by the right hemisphere) such as "she won't do what I tell it," or "I turn the water tap with my right hand and the left comes and turns it off" (Zaidel, 1994, p. 16-17). However, the illusion of unity remains intact. That is, the self-identity is maintained as the left hemisphere always assumes responsibility for the right hemisphere's behaviour (Marinsek, 2016). Patients never say, "my right hemisphere turned the tap off." According to Gazzaniga (2000), it is the left hemisphere interpreter module which helps to maintain this illusion of unity, despite not having access to the modules in the right hemisphere.

Future Endeavors

It is, of course, impossible to propose a theory so ground-breaking that it renders everyone speechless. Many questions surface that of yet have no definitive answer and require future research or clarification. One pressing issue is that Gazzaniga's (2018) version of modularity may not avoid the binding problem—how the cognitive system neurologically integrates information from neuromodular processing into a unified whole. In particular, the binding problem in neuromodular research describes how to scale the activity of a singular neuron into the functioning of a neural network (Revonsuo, 1999), and even further, how such networks work together to create a unified consciousness. These biological details remain unknown.

Conclusion

As I type, many bubbles are rising to bring to me the illusion of a unified consciousness. Audition modules stitch together the dreamy notes of Chopin and the *tap tap* of my keyboard; olfactory modules detect sage with hints of wet dog; visual modules process the dim light from my lampshade, and further, the words from this article. In the

latter case, the neuromodular processing is so fast and instinctual, I *have to* recognize words. Whatever is going on behind the curtain is complicated and vexing—at least for now. I predict future research will uncover that consciousness is a multi-modal system; I also believe that Gazzaniga's (2018) framework for the neuromodularity of consciousness will adapt and mature as a result. Once the architecture is uncovered, we will see the naked truth about consciousness: a sleight of hand of the most phenomenal sort.

References

Albert, M. L., Sparks, R. W., & Helm, N. A. (1973). Melodic intonation therapy for aphasia. *Archives of neurology*, *29*(2), 130-131.

Behrmann, M., Watt, S., Black, S. E., & Barton, J. J. S. (1997). Impaired visual search in patients with unilateral neglect: an oculographic analysis. *Neuropsychologia, 35*(11), 1445-1458.

Beurskens, R., Steinberg, F., Antoniewicz, F., Wolff, W., & Granacher, U. (2016). Neural correlates of dual-task walking: Effects of cognitive versus motor interference in young adults. *Neural plasticity, 2016.*

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE journal on robotics and automation, 2*(1), 14-23.

Conklyn, D., Novak, E., Boissy, A., Bethoux, F., & Chemali, K. (2012). The effects of modified melodic intonation therapy on nonfluent aphasia: A pilot study. *Journal of Speech, Language, and Hearing Research, 55*(5), 1463-1471.

Dennett, D. C. (2003). *The illusion of consciousness* [Video file]. TED. https://www.ted.com/talks/dan_dennett_the_illusion_of_consciousness#t-40336

Doyle, J. C., & Csete, M. (2011). Architecture, constraints, and behavior. *Proceedings of the National Academy of Sciences, 108*(Suppl. 3), 15624-15630.

Durgin, F. H., Tripathy, S. P., & Levi, D. M. (1995). On the filling in of the visual blind spot: Some rules of thumb. *Perception, 24*(7), 827-840.

Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies, 23*(11-12), 11-39.

Flourens, P. (1846). *Phrenology examined*. Hogan & Thompson.

Fodor, J. A. (1983). *The modularity of mind*. MIT press.

Friederici, A. D. (2015). White-matter pathways for speech and language processing. *Handbook of clinical neurology, 129*(3), 177-186.

Gall, F. J., & Spurzheim, J. (1810–1819). *Anatomie et physiologie du système nerveux en général, et du cerveau en particulier.* Schoell.

Gazzaniga, M. S. (1988). Brain modularity: Towards a philosophy of conscious experience. In A. J. Marcel & E. Bisiach (Eds.), *Consciousness in contemporary science*, 218–238. Clarendon Press/Oxford University Press.

Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: does the corpus callosum enable the human condition? *Brain, 123*(7), 1293-1326.

Gazzaniga, M. S. (2018). *The consciousness instinct: Unraveling the mystery of how the brain makes the mind*. Farrar, Straus and Giroux.

Gazzaniga, M. S, Ivry, R. B, & Mangun, G. R. (2014). *Cognitive neuroscience: The biology of the mind. Fourth edition*. W. W. Norton & Company, Inc.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature reviews neuroscience, 8*(5), 393-402.

Jeffries, K. J., Fritz, J. B., & Braun, A. R. (2003). Words in melody: An H215O PET study of brain activation during singing and speaking. *Neuroreport, 14*(5), 749-754.

Kerkhoff, G. (2001). Spatial hemineglect in humans. *Progress in Neurobiology, 63*(1), 27.

Marinsek, N. L. (2016). A split-brain perspective on illusionism. *Journal of Consciousness Studies, 23*(11-12), 149-159.

Merabet, L. B., Hamilton, R., Schlaug, G., Swisher, J. D., Kiriakopoulos, E. T., Pitskel, N. B., Kauffman, T., & Pascual-Leone, A. (2008). Rapid and reversible recruitment of early visual cortex for touch. *PLoS one, 3*(8), e3046.

Nagel, T. (1974). What is it like to be a bat?. *The philosophical review, 83*(4), 435-450.

Nakano, T., Suda, T., Okaie, Y., Moore, M. J., & Vasilakos, A. V. (2014). Molecular communication among biological nanomachines: A layered architecture and research issues. *IEEE transactions on nanobioscience, 13*(3), 169-197.

Nelson, M. E., & Bower, J. M. (1990). Brain maps and parallel computers. *Trends in neurosciences, 13*(10), 403-408.

Overmier, J. B. (2002). Sensitization, conditioning, and learning: Can they help us understand somatization and disability? *Scandinavian journal of psychology, 43*(2), 105-112.

Prinz, J. J. (2006). Is the mind really modular? *Contemporary debates in cognitive science, 14*, 22-36.

Revonsuo, A. (1999). Binding and the phenomenal unity of consciousness. *Consciousness and cognition, 8*(2), 173-185.

Schlaug, G., Norton, A., Marchina, S., Zipse, L., & Wan, C. Y. (2010). From singing to speaking: Facilitating recovery from nonfluent aphasia. *Future neurology, 5*(5), 657-665.

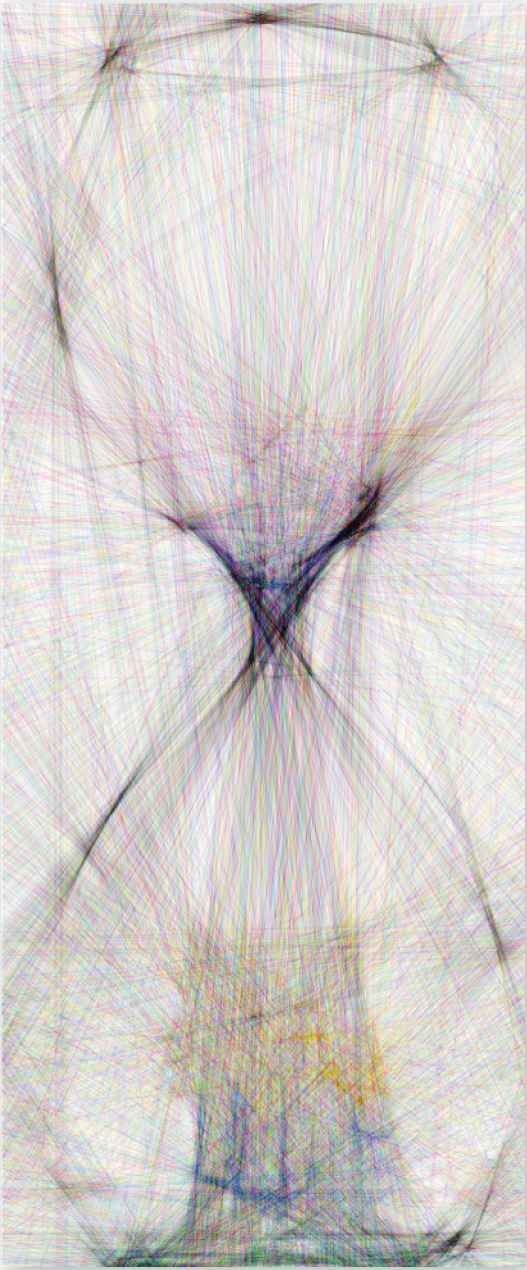
Sperry, R. W. (1969). A modified concept of consciousness. *Psychological review, 76*(6), 532.

Sperry, R. W. (1970). Perception in the absence of the neocortical commissures. *Perception and its Disorders, 48*, 123-138.

Vallar, G. (1998). Spatial hemineglect in humans. *Trends in cognitive sciences, 2*(3), 87-97.

Yamadori, A., Osumi, Y., Masuhara, S., & Okubo, M. (1977). Preservation of singing in Broca's aphasia. *Journal of Neurology, Neurosurgery & Psychiatry, 40*(3), 221-224.

Zaidel, D. W. (1994). A view of the world from a split-brain perspective. In *Neurological Boundaries of Reality, edited by EMR Critchley* (pp. 161-174). Farrand Press.



The Sands of Time Construct, Constrict, Confound the Mind
John Maste

A Representational Account of Lexical Preferences in Quantifier Scope Ambiguity

Jane S.Y. Li
Simon Fraser University

The sentences each student read a book and all students read a book yield different interpretation preferences when read by monolingual English speakers—the former sentence generates a picture of a group of students reading different books, whereas the latter a picture of the same group of students reading a single book (Feiman & Snedeker, 2016). In this article, I propose a psychologically-realized account driven by mental representations of the universal quantifier (Knowlton et al., 2019), which deviates from traditional rule-based quantifier scope theories (Beghelli & Stowell, 1997; Champollion, 2010, and more). Furthermore, the representational theory is able to extend to multilingual data (Scontras et al., 2017) when models of bilingual lexical access (Dylman & Barry, 2018) are considered.

Doubly quantified sentences may generate ambiguous interpretations, as shown in the sentences below:

(1) **Every student read a book.** (∀> ∃)
   a. Surface interpretation (∀> ∃): for all students, they each read a (different) book.
   b. Inverse interpretation (∃ > ∀): a single book was read by all students.

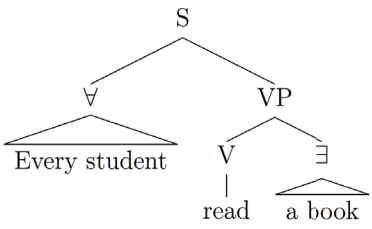
(2) **A student read every book**. (∃ > ∀)
   a. Surface interpretation (∃ > ∀): a single student read all the books.
   b. Inverse interpretation (∀ > ∃): all books were read by (different) students.

The ability to access two logical interpretations from one sentence is often explained by abstract covert movements that occur in between S-structure (where surface forms are represented as trees) and logical form (LF; where logical interpretations may be accessed). In most analyses of quantification-related issues, the operation Quantifier Raising (QR; May, 1990) is assumed. QR specifies that quantifier phrases (e.g. *every student, a book*) may rise to scope over another quantifier phrase at LF. Figure 1 illustrates an example of QR.

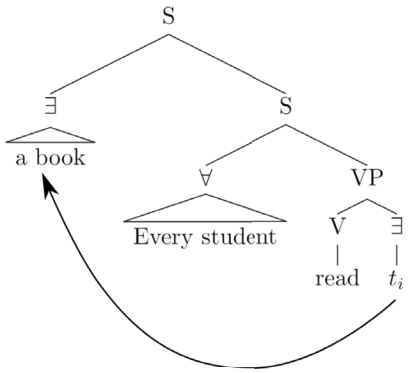
While monolingual English speakers find a preference for the surface interpretation in sentence (1) and (2) (Feiman & Snedeker, 2016; Scontras et al., 2017), the preference changes when the quantifier word is different:

(3) All students read a book.

(4) Each student read a book.



(a) Interpretation: the students read their books separately.



(b) Interpretation: A single book was read by all the students.

Figure 1. Derivations of the surface and inverse interpretations via QR.

The usage of *all* in sentence (3) induces a strong preference for the inverse interpretation (a single book was read), whereas *each* induces the preference for the surface interpretation (for all students, they read a different book). Despite the logical equivalence (in surface form) in both first and second-order logic, the three quantifier words *each, every,* and *all* lead to robustly different interpretation preferences. More specifically, Ioup (1975) observed that quantifier words fall along a hierarchy based on their preference for the surface interpretation:

each > every > all > most > many > several > some > a few

Many works on quantifier scope ambiguity have set out to provide an adequate explanation of lexical preferences in quantifier scope ambiguity (Beghelli & Stowell, 1997; Champollion, 2010), with most of them postulating an abstract feature that dictates scoping mechanisms that give rise to inverse and surface readings. Is there, however, a less abstract explanation that accounts for the graded hierarchy that has been frequently replicated?

Furthermore, a much more complex pattern of scopal preferences arises when we look beyond monolingual English data. Scontras et al.'s (2017) investigation of bilingual scope access made use of heritage Mandarin speakers, who speak both Mandarin (L1; an inverse-prohibiting language, where inverse readings are strongly dispreferred) and English (L2; inverse-allowing), but have used English dominantly since the onset of schooling. When tested on surface versus inverse preferences in English doubly quantified sentences, the heritage speakers patterned closer to monolingual Mandarin speakers than monolingual English speakers (but significantly different from both groups) despite being more proficient in English.

The prohibition of inverse scope in Mandarin is a well-documented phenomenon (Aoun & Li, 1989; Huang, 1998) which has been accounted for by the Isomorphic Principle (Aoun & Li, 1989):

(5) Suppose A and B are Quantifier Phrases. Then if A c-commands B at S-Structure, A c-commands B at LF.

In other words, if a quantifier phrase such as *every student* scopes over another quantifier phrase (*a book*) in its surface form (Figure 2), it must remain that way in the interpretation, such that only surface interpretation can be accessed.

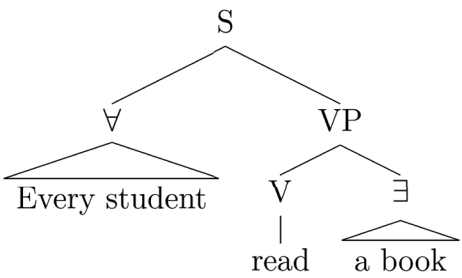


Figure 2. Every student scopes over (c-commands) *a book*

From their experimental results, Scontras et al. (2017) conclude that the Isomorphic Principle applies to the system of the heritage speakers (i.e. it applies to all languages that the person speaks) because it is the simpler configuration. Subsequently, Scontras et al. (2017) claims that the deviation of heritage speakers from the monolingual Mandarin speakers stems from a well-documented "yes-bias" (Scontras et al., 2017, p. 19) that enables heritage speakers to easily accept an interpretation rather than reject it. Again, like the explanations for monolingual English data, is there a more naturalistic explanation that takes into account multilingual data without resorting to experimental design-related biases?

In sum, a successful account of quantificational preferences and access must provide an adequate explanation for lexical preferences across quantifier words. In this paper, I propose a representational theory that draws inspiration from experimental work on the representation of universal quantifiers (Knowlton et al., 2019) and lexical access models in psycholinguistics (Dell, 1986; Dylman & Barry, 2018). Instead of postulating a feature which dictates lexical preference, I propose that lexical preferences are driven by the order of logic (i.e. propositional/predicate vs. set theoretic representations) of each quantifier word, which is consistent with the experimental findings from Feiman and Snedeker (2016). I also posit that the theory can be extended to multilingual data if we consider spreading activation of prohibiting features from their heritage language to their target language (English). In conjunction with other non-conflicting principles, the representational theory provides a psychologically-grounded explanation for lexical preferences in scope ambiguity.

In the following sections, I will first elaborate in finer detail a traditional analysis of lexical preferences in the context of current empirical data. In turn, the shortcomings of this framework motivates a new analysis that deviates from traditional rule-based theories. In section three, the representational framework is discussed in detail—from its theoretical basis to its implications for future work, this paper provides an in-depth analysis of scopal semantics by incorporating psycholinguistic findings and linguistic theory.

The Traditional Analysis

This section is first prefaced with a discussion on preference and accessibility, which is crucial for interpreting experimental results and

creating theoretical inferences. Then, the analyses proposed in previous literature shall be examined in detail in light of empirical results reviewed in the previous section.

Preference and Accessibility

Literature relevant to quantifier scope ambiguity use the terms *preference* and *accessibility* as a descriptor of one's response towards interpretations of quantificational ambiguity.

The notion of **accessibility** is generally used to describe the **existence** of an interpretation in the minds of the reader. For instance, the inverse interpretation of sentence *every student read a book* is accessible for English speakers, but not for Mandarin speakers (Aoun & Li, 1989; Huang, 1998)—when prompted with a question, "can all the students be reading a single book together?" (inverse), most Mandarin speakers would answer "no", as they lack the accessibility for the inverse interpretation. Accessibility can be thought of as a binary feature applied to an interpretation, where a reading is either [+accessible] or [−accessible]. In an acceptability experiment, where subjects are asked to rate a certain interpretation on a scale of 0 to *n*, a "floor" rating (0–1 on a scale of 7) is generally reflective of an inaccessible interpretation.

Preference, on the other hand, is concerned with the **strength** of the interpretation. When prompted with the same question, "can all the students be reading a single book together?", most English speakers would answer along the lines of, "Yes, but it makes more sense when they're each reading their own books" (surface), as they have a stronger preference for the surface interpretation. As doubly quantified sentences come with only two possible interpretations, preference can be thought of as relative and binary—when one interpretation is preferred, it implies that it is preferred over the other interpretation, and the other interpretation is **dispreferred.** The magnitude of preference/dispreference can be thought of as being on a continuum, with strong on one end and equal preference on the other end. On the measure of acceptability, the strength of preference is generally reflected by the magnitude of acceptability difference across conditions (e.g. surface versus inverse).

The Distributivity Analysis

We now turn to a classic analysis in the generative fashion, presented by Beghelli and Stowell (1997). It is important to note that Beghelli and Stowell's (1997) framework is designed to account for a

host of quantification-related phenomena, thus parts of the analysis here might seem superfluous when presented in this article.

Beghelli and Stowell (1997) first brings up the notion of *distributivity*—a quantifier word is distributive when it addresses the referents individually, as opposed to non-distributive quantifier words which address the referent as a single unit. Distributivity can be diagnosed in the following sample sentences:

(6) a. ?? **All** students read a different book.
b. **Each** student read a different book.

In sentence (6), a surface interpretation is necessary to correctly understand the sentence, where books are read individually by each student. This set of sample sentence diagnoses *each* as a distributive quantifier

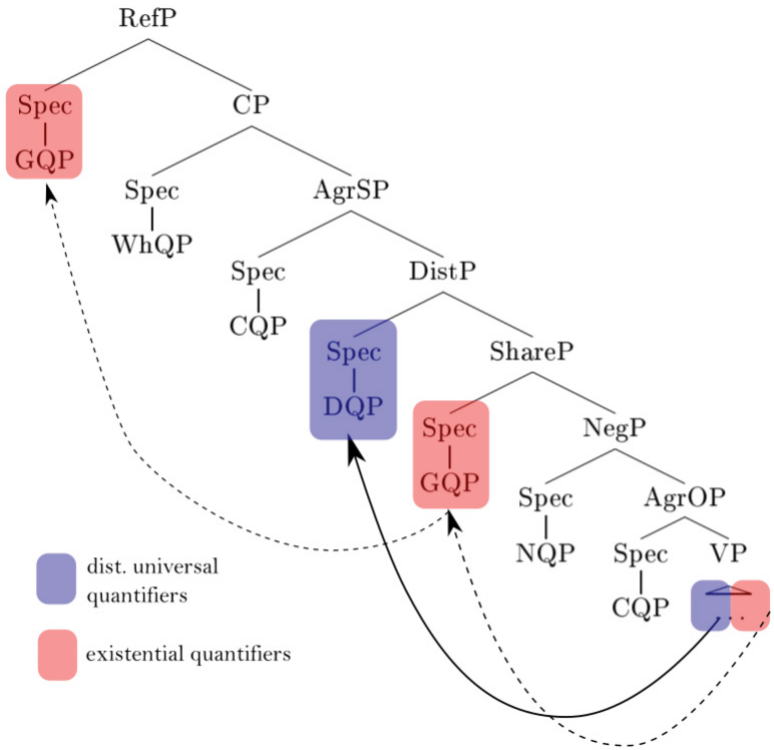


Figure 3.

Note. LF movement of [+dist] quantifier words, where the quantifier phrase necessarily c-commands the existential quantifier phrase. Tree taken from Beghelli and Stowell (1997).

word, and *all* non-distributive. The authors then postulate that distributivity is a syntactic feature, and establish that *each* is [+distributive], *every* is unspecified, and *all* is [–distributive].

Then, in the context of covert movement at LF, Beghelli and Stowell (1997) propose that [+distributive] words undergo obligatory movement to a distributive quantifier phrase (DQP) in order to get its feature checked at [Spec, DistP] (Figure 3).

Conversely, [–dist] quantifier words are disallowed from movement to [Spec, DistP], thus predicted to not have access to the surface interpretation in the sentence *all students read a book*. Unspecified [dist] words (*every*) have optional movement. The [dist] specifications on the three universal quantifier words resemble Ioup's (1975) hierarchy (each > every > all).

Group-denoting quantifier phrases (essentially existential quantifier phrases) may undergo raising to scope over DQPs in order to generate the inverse interpretation.

In sum, Beghelli and Stowell's (1997) analysis postulates an abstract feature [±distributive] that dictates whether or not universal quantifier phrases are allowed to raise to scope over the existential quantifier phrases. We now turn to a representational account that attempts to account for the same set of monolingual data.

A Representational Account

As concluded by Feiman and Snedeker (2016), "[f]uture linguistic theories should seek to explain the representational basis of these preferences, and how they relate to the mechanisms underlying quantifier scope assignment and LF construction" (Feiman & Snedeker, 2016, p. 50). This section echoes Feiman and Snedeker's vision for theories of quantifier scope ambiguity. I propose a new framework built on previous investigations of the mental representations of linguistic units (Knowlton et al., 2019), which can successfully predict lexical preferences for monolingual English speakers. Furthermore, with the consideration of bilingual models of lexical access, the representational theory may extend to account for bilingual data elicited from Scontras et al. (2017).

Basic Assumptions

Like many theories of quantifier scope ambiguity, this thery assumes that Quantifier Raising (hereafter QR) is the main scoping mechanism to derive inverse interpretations. I also follow current

theories in postulating that the Processing Scope Economy Principle (Anderson, 2004) and Isomorphic Principle (Aoun & Li, 1989, see section 1, (5)) hold true:

(7) The human sentence processing mechanism prefers to compute a scope configuration with the simplest syntactic representation (or derivation). Computing a more complex configuration is possible but incurs a processing cost. (Anderson, 2004)

More precisely, the Isomorphic Principle is only applied to inverse-prohibiting languages, and not to the language user's system like Scontras et al. (2017) have suggested. Without postulating other mechanisms, these principles predict that when one is using an inverse-prohibiting language, they are unable to perform QR and access the inverse interpretation. Once they speak an inverse-allowing language, they have access to the inverse reading, but it is dispreferred due to its higher processing cost.

Mental Representation of Universal Quantifiers

In the realm of classical logic, the following predicates are logically equivalent but are represented in different order:

(8) Every book is heavy.
a. **First-order logic** (predicate and propositional):
$\forall_x B(x) \Rightarrow H(x)$. (for all entities that are books, they are heavy.)
b. **Second-order logic** (set-theoretic):
$B \subseteq H$, where $B = \{x : x \text{ is a book}\}$ and $H = \{x : x \text{ is heavy}\}$.
(the set of all books are contained within the set of heavy things.)

First-order logic terms embody a relationship between properties (as predicates) of an abstract entity, whereas second-order terms embody a relationship between sets of entities.

Through verification tasks, Knowlton et al. (2019) probed the mental representation of the universal quantifier words: *each*, *every*, and *all*. Subjects were first shown quantified sentences such as *all of the/every/each of the big dots are blue*, then a visual display related to the initial sentence. They were then given a question regarding cardinality of the visual scene shown earlier, and the accuracy and precision of their answers were the dependent variables of the study.

They hypothesize that if a sentence is mentally represented in second-order form, subjects are more sensitive toward numerical information, such as the cardinality and the amount of overlapping members between two sets, thus performing more accurately and precisely. *Most* was used as baseline to confirm this hypothesis as the access to *most* must involve the representation of set and cardinality information:

$$\text{Most A are B iff } |A \cap B| > |A \cup B|$$

The number of members in the set of items that are A and B is larger than the set of things that are A and not B.

Results from the verification tasks suggest that *every* and *all* pattern with second-order logic, whereas *each* patterns with a first-order terms. More specifically, however, Knowlton et al. (2019) found that responses for *all* target sentences were significantly more precise than *every* sentences—*all* holds a stronger second order representation with respect to precision. On that view, precision across universal quantifier words present an *each* > *every* > *all* hierarchy.

Instead of postulating a [± distributive] feature that dictates interpretation preferences, I propose that the order of logical system (first vs. second) is the key context for the divide in scoping behaviour—universal quantifier words that pattern with first order logic (*each*) prefer surface readings; conversely, second order logic words (*all*) prefer inverse readings. I speculate that inverse constructions via quantifier raising with sets as an entity (second order) are processing-wise, less costly than their predicate counterparts, where each item in the observing set are seen as individuals. This speculation is somewhat similar to Beghelli and Stowell's (1997) reasoning, but in the context of processing cost as opposed to obligatoriness in covert movement. Not only is this postulation is consistent with Feiman and Snedeker's (2016) findings for universal quantifiers, it also provides a naturalistic reason for the difference in scoping behaviour. The following example illustrates how lexical preferences can be derived in the current framework.

Given two doubly quantified sentences *each student read a book* (4b, 5b) and *all students read a book* (4a, 5a), their S-structures can be represented by the following trees:

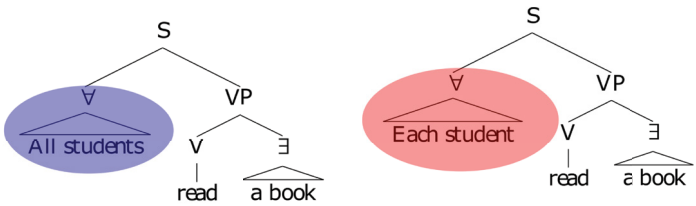


Figure 4 . The two doubly quantified sentences at the S-structure level.

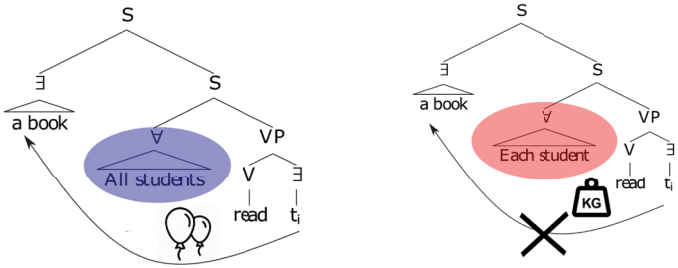


Figure 5. The two doubly quantified sentences at logical form.

As noted in the basic assumptions of the theory, inverse interpretations shall be derived through quantifier raising. The theory predicts that the inverse processing cost is eased for *all students read a book* given the strong second-order representation of *all*. Conversely, the inverse processing cost for *each student read a book* is increased due to the strong first-order representation of *each*.

Spreading Activation of Prohibitive Features

Thus far, we have taken into account data for monolingual speakers. Now, we turn to the bilingual data, particularly that of heritage speakers of Mandarin. In this section, I entertain the possibility of using connectionist models to account for bilingual data. It is important to note that experimentation, in this case computational modelling is required to verify my claims in this section.

Theories of bilingual lexical access exist on a fairly wide spectrum (Rapp & Goldrick, 2006), but scholars in that domain have a general consensus on several properties of the lexical access model, such as connections between translation equivalents. For the purposes of this paper, I will be taking these properties as granted.

Spreading activation models, such as Dell's (1986) postulation that linguistic units are stored as nodes and get activated to produce utter-

ances (Figure 6). The activation spreads top-to-bottom (i.e. syntax to phonology) and the link between nodes are weighted according to their relevance (e.g. *swimmer* activates the morphemes *swim* and *-er* strongly, but activation also spreads to other nearby nodes since connections exist). Dell (1986) proposed this model to initially account for speech errors in English, but the model has quickly extended in various directions to account for other linguistic phenomena.

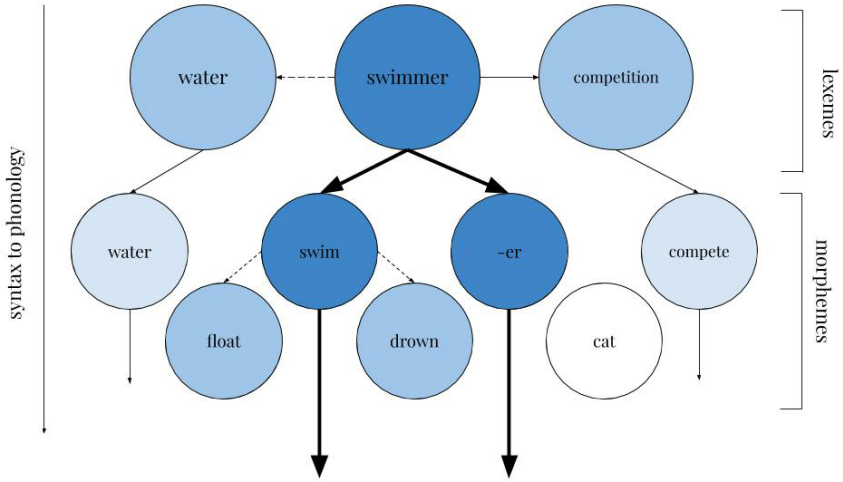


Figure 6. A visualization of the spreading activation model from Dell (1986).
Note. The shade of blue indicates the level of activation and the thickness of the arrows indicates the connection strength between nodes.

Many of these efforts have been applied to bilingual interaction—bilinguals often display transfer, where they apply features of one language to the use of the other (Gass, 1984; Poulisse & Bongaerts, 1994). Scholars have proposed that this can also be accounted for by spreading activation models, where features of one language spread to other language, resulting in similar behaviours as speech errors, but in a multilingual fashion. For example, (Dylman & Barry, 2018) propose an account of bilingual lexical storage, where lexical items, when activated, facilitate activation of their lexical-equivalent in the other language.

I propose that, like other transfers, scoping mechanisms of heritage Mandarin speakers can also be accounted for by the spreading activation between Mandarin and English nodes. By the Isomorphic

Principle, quantifier raising only applies to the Mandarin nodes but not the English nodes. When acceptability ratings of English sentences are elicited from heritage speakers, the activation of the Mandarin nodes spread its prohibiting features to the target English nodes; conversely, the inverse-allowing properties of the English nodes spread to the Mandarin nodes when Mandarin judgements are elicited. This assumption successfully predicts the graded behaviour (not as prohibiting as Mandarin, but not as accepting as English) of heritage Mandarin speakers toward English and Mandarin doubly quantified sentences, without resorting to interpretations such as yes-bias.

Discussion, Limitations, and Future Work

The new proposal sheds light on how lexical preferences in scope ambiguity may be derived through naturalistic means. This comes with a host of implications, both on the issue of scope ambiguity and linguistic theory as a whole.

While my proposed theories tackle a subset of the issues in quantifier scope ambiguity, the tools and mental representations reported here might be deemed useful for other issues in scope ambiguity. For example, the issues of negation and its scope may benefit from an analysis through connectionist networks, particularly when considering cross-linguistic perspectives. On a larger scale, I have demonstrated there is a psycholinguistic place for what is traditionally thought as a theoretical linguistics problem. My hope here, is for this work to initiate a conversation about stronger incorporation of psycholinguistic and theoretical linguistic theories, especially at the level of the syntax-semantics interface.

Limitations

First, most of the experimental findings reviewed in this paper have been focused on universal quantifiers and their psychological nuances. Therefore, most of the argumentation made in this paper are based on universal quantifiers, and hypothesized to be able to extend to existential quantifiers. Future experimentation on existential quantifier is recommended—an integration of Knowlton et al. (2019) and Scontras et al.'s (2017) elicitation methods would be optimal in extracting the mental representation of existential forms.

Second, the interactions between universal/existential quantifiers and negation were not discussed. The previous analyses on preferences in scope ambiguity have set out to model other related phenomena

such as negation interaction (e.g. Beghelli and Stowell (1997))—for a fuller view of scope ambiguity, other biases in scope ambiguity should also be surveyed in my representational theory.

Third, most of the experimental work that I have reviewed for this article is fairly recent, thus their results and implications are yet to be replicated and supported by other experimental evidence. Therefore, the relevancy of this paper is dependent on the replicablity of the experimentation and theories cited in this paper.

Conclusion

In this paper, a new analysis of quantifier scope ambiguity was proposed—it takes into account the graded behaviour of scopal preferences across quantifier words, as well as the interaction between multiple languages in a system. The representational account deviates strongly from previous rule-based accounts, which generally postulate abstract categorical features in attempt to account for graded data.

Looking beyond the theory as an explanation for the data at hand, I hope that this article inspires future works that incorporate psychologically-realized explanations. Not only does this encourage interaction between the subfields of cognitive science, it also pushes the bounds of linguistic inquiry.

References

Anderson, C. (2004). *The structure and real-time comprehension of quantifier scope ambiguity* (Doctoral dissertation). Northwestern University Evanston, IL. Aoun, J. & Li, Y.-h. A. (1989). Scope and constituency. Linguistic inquiry, 20 (2), 141–172.

Barker, C. (2002). Continuations and the nature of quantification. *Natural language semantics*, 10 (3), 211–242. Beghelli, F. & Stowell, T. (1997). Distributivity and negation: The syntax of each and every, In Ways of scope taking. Springer.

Beghelli, F. & Stowell, T. (1997). Distributivity and negation: The syntax of each and every, In *Ways of scope taking*. Springer.

Benmamoun, E., Montrul, S. & Polinsky, M. (2013). Heritage languages and their speakers: Opportunities and challenges for linguistics. *Theoretical Linguistics,* 39 (3-4), 129–181.

Champollion, L. (2010). *Parts of a whole: Distributivity as a bridge between aspect and measurement* (Doctoral dissertation). University of Pennsylvania.

Cooper, R. (2013). *Quantification and syntactic theory* (Vol. 21). Springer Science & Business Media. Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. Psychological review, 93 (3), 283.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review, 93* (3), 283.

Dylman, A. S. & Barry, C. (2018). When having two names facilitates lexical selection: Similar results in the picture-word task from translation distractors in bilinguals and synonym distractors in monolinguals. *Cognition*, 171, 151–171.

Feiman, R. & Snedeker, J. (2016). The logic in language: How all quantifiers are alike, but each quantifier is different. *Cognitive Psychology,* 87, 29–52.

Fox, D. (1995). Economy and scope. *Natural language semantics, 3*(3), 283–341.

Gass, S. (1984). A review of interlanguage syntax: Language transfer and language universals. *Language Learning, 34* (2), 115–132.

Huang, C.-T. J. (1998). *Logical relations in chinese and the theory of grammar.* Taylor & Francis.

Ioup, G. (1975). Some universals for quantifier scope. *Syntax and semantics, 4,* 37–58.

Knowlton, T., Pietroski, P., Halberda, J. & Lidz, J. (2019). The mental representation of universal quantifiers: Evidence from verification. *Manuscript in preperation.*

May, R. C. (1990). *The grammar of quantification.* Garland.

Montague, R. (1973). The proper treatment of quantification in ordinary english, *In Approaches to natural language*. Springer.

Pezzelle, S., Bernardi, R. & Piazza, M. (2018). Probing the mental representation of quantifiers. *Cognition, 181*, 117–126.

Poulisse, N. & Bongaerts, T. (1994). First language use in second language production. *Applied linguistics, 15* (1), 36–57.

Rapp, B. & Goldrick, M. (2006). Speaking words: Contributions of cognitive neuropsychological research. *Cognitive Neuropsychology, 23* (1), 39–73.

Rumelhart, D. E. & McClelland, J. L. (1986). On learning the past tenses of english verbs. *Parallel distributed processing.*

Scontras, G., Polinsky, M., Tsai, C.-Y. E. & Mai, K. (2017). Cross-linguistic scope ambiguity: When two systems meet. *Glossa: a journal of general linguistics, 2* (1), 1–28.

Brain Powered
Deirdre Harder

Be Wise and Envy Free: Investigating Coping Strategies of Malicious Envy

You Zhi Hu
University of Toronto

Malicious envy is often examined as one form of self-destructive behaviour—compulsive and resistant to change. Malicious envy has four main mechanisms: misattribution of salience, inner conflict, weakness of will, and unconscious difference monitoring. This paper develops its argument by 1) summarizing the description of target core mechanisms of malicious envy, 2) examining the currently available coping strategies to tackle those mechanisms, and 3) arguing that the effectiveness of the strategies is tightly connected with reframing to combat malicious envy. Reframing is a universal mechanism shared by all coping strategies corresponding to the four mechanisms of malicious envy, which helps one re-evaluate situations and values of envied targets in order to support one to look into reality and be free from envy.

Malicious envy is greed and jealousy that seeks to possess and desire something good with the intent to destroy the desired objects (Rathbone, 2012). It is characterized as a feeling of deficiency caused by others' superiority and hostility demonstrated in a desire to pull down the envied others (Ven et al., 2009). For this reason, it is considered as one form of self-destructive behaviour, which leads to poor mental and physical health with the experience of frustration and resentment (Smith et al., 2008).

There are several popular hypotheses of the origin of malicious envy. For example, malicious envy is often said to be caused by upward social comparisons (Ven et al., 2011). When one makes an upward comparison from a low hierarchy position, self-devaluation and inferiority occur (Wheeler, 1966). Exaggerating the importance of other

people's possessions leads to negative social cognition. This is the first mechanism of malicious envy—misattribution of salience. On some occasions, the agent's own inner conflict between the ideal self and actual self misdirects hostility towards others (Rogers, 1959), which is the second mechanism of malicious envy. In simpler words, malicious envy could be a form of blaming others for one's own underperformance. Thirdly, envy arises from weakness of will, when one becomes incapable of controlling one's destructive thoughts and behaviours to achieve well-being (Davidson, 2001). Lastly, unconscious difference monitoring serves an evolutionary purpose that allows humans to adapt to the environment (Gerhardt, 2009). However, this survival mechanism also makes individuals negatively evaluate their own self-perception against their evaluation of others, which eventually results in a sense of injustice, low self-esteem and resentment (Smith, 1991). Through integrating the literature, this paper summarizes these four core mechanisms of malicious envy: misattribution of salience, inner conflict, weakness of will and unconscious difference monitoring.

In addition, this paper also proposes coping strategies for each mechanism by examining theories and empirical data. In the field of psychological counselling, Cognitive Behavioral Therapy and mindfulness practice are commonly used practices for treating malicious envy to mediate misattribution of salience and alleviate inner conflict (Beck, 1995; Tabak, 2015). Strategies include building self-reliance and a technique called "depreciating values of resources" which can counter weakness of will and unconscious difference monitoring (Quintanilla & Jensen de Lopez, 2013). However, the effectiveness and mechanisms of these treatment/coping strategies still need to be validated. Ultimately, the fundamental mechanism of these coping strategies is *reframing*, which shifts one's attention to reassess the current situation with relevant information and reformulates the value of the envied target to resolve irrationality (Palumbo, 2004). Thus, reframing allows people to re-evaluate what they see and to mitigate the effects of envy.

Misattribution of Salience

Misattribution of salience is a fundamental contributor to malicious envy. Misattribution of salience means attributing importance to non-important objects, which brings non-salient objects to the foreground to be focused on and paid attention to. Notably, in malicious envy, subjects attribute importance to others' possessions, seeking to

take over the possessions with the intent to destroy them (Rathbone, 2012). Therefore, people with malicious envy tend to deem arbitrary items as valuable and desirable (e.g., the neighbour's TV), and are preoccupied with such desire. One of the most effective treatments for malicious envy is **Cognitive Behavioural Therapy (CBT)**. CBT generally produces a cognitive modification in patients' thinking and belief system to solve a broad range of psychological disorders and problems (Beck, 1995). It is effective in treating depression, anxiety and borderline personality disorder (Davidson, 2006). Research has suggested that CBT can mitigate the negative feeling of envy. In a 2017 study, participants received seven weekly therapy sessions with CBT, followed by interviews and thematic analysis. Results demonstrate increased insights and a positive shift in emotion. Particularly, they observed significantly less intense envy and improved self-confidence (Cotter et al., 2017). Participants reported that they felt happier and noticed an overall improvement in their attitude. In addition, CBT also helped participants become more aware of their thoughts and feelings. In CBT sessions, patients were guided to detach from the previously established value system. The patients were then able to objectively re-evaluate the envy-provoking situation. Finally, when the patients became aware of their thoughts and consciousness, they were able to identify the disproportional values they have attributed to the envied item (e.g., having a fancy TV is actually not important).

The effectiveness of CBT is mainly due to its insight-evoking procedure. The insight phenomenon is when an old thought pattern is broken, relieving cognitive constraints, and new ways of construing information are formed (Vervaeke et al., 2012). Insight, as a process of replacing bad problem formulation with good problem formation, is generated from reframing. Reframing formulates a re-evaluation of the current situation (Palumbo, 2004). It is an essential strategy to directly cope with misattribution of importance. In solving malicious envy, we need to have the insight to shift the fundamental framing to reassess and reformulate the value of the envied target. Harris and Salovey (2008) suggested "reframing envy-provoking situations" as a coping strategy in their reflection on envy; "reassessing desires, beliefs and goals" helps people realize that the envied item is not so important and also help them clarify future goals to achieve. This is to say, reframing the envy-provoking situation makes real importance salient to people.

Reframing by reformulating the problem brings more salient things to the foreground in the real world. Reframing helps people become wise to critically examine what is more salient in the inner state and more important goals in the future.

In addition, I would like to argue that envy is very similar to bullshit. According to Frankfurt's definition of bullshit, it entails an absent direct concern for the truth (Pennycook et al., 2015). Unlike lying, in which the speaker knows the truth and tries to hide it, people who bullshit do not care about the truth and only care about whether their listener is persuaded (Frankfurt, 2005). When bullshitting, people unconsciously make a particular statement more salient than it actually is. In parallel, when envying others, people believe others' possessions are absolutely essential and relevant to take over; people unconsciously misidentify the value of their attributed importance and believe the desired targets are more salient. Thus, malicious envy works in the same way as bullshit, which misattributes salience to envied targets. Furthermore, reframing can resolve both envy and bullshit. To resolve both of these phenomena, one needs to examine and re-establish the truth value of the associated statement. Hence, reframing is a universal mechanism of coping strategies to resolve irrationality and self-destructive behaviour, including bullshit and envy. It can detach the truth value assigned to bullshit or envy and reattribute an appropriate and vital relevance instead of having illusions of others having more essential possessions.

Inner Conflict

Inner conflict is another mechanism pertinent to envy. In traditional psychodynamic theory, Freud proposes three agents in the structural model of the psyche: 1) the ego, which is the mediator between 2) the id and 3) the superego (Freud & Strachey, 1966). Carl Rogers also suggests that inner conflict arises from an incongruence of self-concept between an actual self and an ideal self, and the experience primarily causes anxiety (Rogers, 1959; Sato, 2005). Self-concept comprises an ideal self and an actual self. This schema is analogous to Freud's psychodynamic theory, the superego tries to pursue an ideal self, and the id is concerned with an actual self involving desires and instincts. Thus, ego resolving the discrepancy between the id and the superego is similar to mediating inner conflict between an actual and an ideal self. Envy arises from the inability to mediate inner conflict

between the id and the superego or incongruence between the ideal and actual self. In this situation, inner conflict dysregulates emotions. Gerhardt (2009) highlights the fundamental problem of envy as the "unbearable inner realities" of conflict between self and others. In other words, envy emerges from inner conflict due to the dysregulations of the ego with self-concept. Marques illustrates the relationship between the ego and envy: the ego can be tamed by spiritual evolution, which refers to the growth of consciousness (Hawkins, 2001; Marques, 2011). However, if there is no effort toward spiritual evolution, then the ego will continuously result in different types of self-destructive behaviours such as envy, jealousy, and greed (Gerhardt, 2012).

I would like to argue that a better representation of inner conflict is unclear thinking due to a second level of misattribution of salience. Envy occurs when people become unable to attribute what is more salient. In the previous section, this paper addresses misattribution of salience, a mechanism that people categorize others' possessions as more important than their own and worth envying. Furthermore, there is a second level of misattribution of salience resulting from inner conflict; the ego misattributes inner conflict (the gap between the ideal self and the actual self) to others' fault, which deepens the tension between the ideal self and the real self. For example, some people may attribute not having a higher income to others being too competitive instead of self-reflecting on how to achieve goals. Therefore, misattribution of inner conflict prevents one from pursuing positive self-enhancement and well-being. Inner conflict is as crucial as misattribution of salience, and neither of them can be discounted in causing envy. Both are essential mechanisms behind envy.

Freud's models, of course, are abstractions, which have been contested over the past century. Nevertheless, neurobiological evidence suggests the existence of inner conflict and its relationship to self-concept. Petchkovsky et al. (2010) gave subjects Jung's 100 Word Association Test, which intended to invoke a state of inner conflict, and then monitored brain activity using functional magnetic resonance imaging (fMRI). They observed a pattern of response in brain regions which include the anterior insula, medial prefrontal, lateral prefrontal, and mid-temporal regions (Petchkovsky et al., 2010). Interestingly, these regions are also associated with the experience of "self" and the discrepancy between the ideal self and actual self (self-con-

cept). Firstly, significant neural activation in the anterior insula is uniquely associated with self-reflection and highly salient information relative to one's sense of self (Modinos, Ormel & Aleman, 2009; Perini et al., 2018). Extrapolating from the above studies, one can infer that when inner conflict occurs, the anterior insula is unable to properly process self-relevant and salient information causing unclear thinking due to misattribution of salience. Another neuroimaging study asked participants to perform self-referent tasks under fMRI by making trait adjective judgement about themselves relative to personally known others (Heatherton et al., 2006). The result shows a relatively robust blood-oxygen-level-dependent (BOLD) signal in the medial prefrontal cortex (mPFC) when participants were making judgements about themselves. Moreover, they conclude that a response from the mPFC is self-specific, which means "judgements pertaining to oneself are distinct from those made for one's friend." The mPFC, as one of the brain regions active for both inner conflict and self-specific concept, entails that malfunctions in self-concept trigger inner conflict introduced by Freud and Rogers. Therefore, neurological evidence suggests that the presence of inner conflict in our brain arises from the discrepancy of self-concept, which leads to unclear thinking and induces envy.

To tackle this inner conflict of envy, **mindfulness practice** has been shown to be effective. Mindfulness practice generally brings people to practice shifting attention to experience the present moment, which is beneficial for psychological health, reducing negative symptoms from psychological disorders like schizophrenia (Tabak, 2015). In coping with envy, mindfulness practice minimizes inner conflict between self and ego by realizing the existence of our biases, other internal desires, and defensive tendencies (Gerhardt, 2012). It helps people obtain and reattribute acute awareness, attentiveness, and perceptiveness in everything from the surrounding environment. An example of mindfulness practice is Appreciative Joy Meditation (AJM). AJM is a practice to appreciate happiness and celebrate the happiness of others by breaking down the barrier of unhappiness. In a 2017 study, AJM was shown to reduce participants' feelings of envy by enhancing other-focused positive emotions (Zeng et al., 2017). In other words, AJM, by regulating emotions in a positive way, alleviates inner conflict, which is caused by the ego's inability to meditate on mental conflict. Additionally, AJM, shifting participants' attention towards the appreci-

ation of joy, mitigates the feeling of envy. The research on AJM suggests that reframing is the fundamental mechanism of coping with envy.

Furthermore, the lateral prefrontal cortex (LPFC) is a region which is not only relative to inner conflict, but also relative to self-control and the reward system. A group of researchers used low-frequency repetitive transcranial magnetic stimulation (rTMS) to disrupt the lateral prefrontal cortex's function (Figner et al., 2010). Then they gave participants a decision-making task, which required a trade-off between short-term and long-term consequences. Result shows that rTMS negatively impacts the reward system with a self-control mechanism. Namely, participants chose the immediate rewards without considering the long-term consequences. Inner conflict inducing envy can be understood as inner conflict negatively impacting the reward system in LPFC, which causes failures in preventing oneself from pursuing immediate rewards. People become incapable of making rational decisions for well-being and ignore the long-term negative consequences of what envy brings to the brain. This research also reminds of a potential solution to resolve the inner conflict, which is to improve the ability of self-control and build up a robust self-regulatory system.

In order to resolve inner conflict, building up a self-regulatory system is the key to be capable of reframing and transcending ego into well-being. Mindfulness may help people form their own self-regulatory systems, which increases people's flexibility in the transparency-opacity shift as shifting attention from looking through to looking at the present conscious state (Metzinger, 2003); this is an introspective grasp to wake the presence of consciousness. Practicing the transparency-opacity shift helps people *look at* the current mental state with the envy-provoking situation rather than *look through* situations with a lens of malicious envy. When people are making decisions *through* inner conflict, this regulatory system puts inner conflict under focal awareness, looking *at* the conflict, to examine and re-evaluate. Once one forms the self-regulatory system, one can critically see things with less bias from the ego, which causes inner conflict. Then one is able to make proper judgements to look through illusions into reality. Therefore, constantly practicing meditation can reduce envy by forming the self-regulatory system and developing reframing to improve well-being.

Weakness of Will

In the previous section, the self-regulatory system is discussed as one way to resolve inner conflict to reduce envy. This self-regulatory system points out another mechanism causing envy, which is weakness of will. The ability to self-regulate is to control one's thoughts, emotions, or behaviours by overriding one's impulses (Baumeister, Heatherton & Tice, 1994; Evans, Boggero & Segerstrom, 2016). Thus, people committing irrational actions that are contrary to their better judgement due to lack of will power are called "weak-willed" (Davidson, 2001). In Greek, such "weak-will" is called *Akrasia*, which means lacking command and acting against one's better judgment. This mechanism can potentially explain the phenomenon of envy. A group of researchers studied the relationship between the phenomenological level of *Akrasia* and its neuropsychological level in the human brain (Kalis et al., 2008). Specifically, they think *Akrasia* is a deficit in executive control of the decision-making process, which highly correlates with the dorsolateral, frontopolar, and orbitofrontal prefrontal cortex as well as the caudate, ventral striatum, anterior cingulate and putamen. Indeed, *Akrasia*, a lack of self-control due to the deficit in executive control, is parallel to envy in its incapability of controlling self-destructive thoughts and behaviours. Aristotle developed the concept of *Enkrateia* as the power of self-control, to contrast with *Akrasia* (Pritchard et al., 1945). A person with *Enkrateia* is capable of obtaining self-control to overcome self-destructive behaviour. Holton argues that such power of self-control is "something like a muscle" that can be developed through hard work and exercise: "the more often agents succeed in exerting self-control, the easier it becomes to maintain their resolutions in the future" (Holton, 2003).

Salovey and Rodin (1988) identified three potential coping strategies for envy: **self-reliance, self-bolstering**, and **selective ignoring**. Self-reliance is avoiding emotional outbursts and maintaining activities without asking others for help. One with self-reliance is able to suppress the degree of envy on one's own without any external intervention. Self-bolstering induces subjects to take everything positively; in this case, practices such as thinking one's good qualities and taking care of oneself mitigate the negative feeling of envy. Selective ignoring tries to ignore the importance of the envied goals. The empirical investigation demonstrated that self-reliance was the most effective

strategy (Salovey & Rodin, 1988). Based on the regression model, self-reliance exhibited the highest negative coefficient, meaning it was most associated with reduced envy. Neither self-bolstering nor selective ignoring was significantly associated with reduced envy (Salovey & Rodin, 1988).

Salovey and Rodin (1988) believe that the reason behind self-reliance's effectiveness is that it reframes the feeling of envy, leading people to re-evaluate the importance of envied targets. Again, this evokes the previous notion that misattribution of importance is the fundamental problem of envy. However, this does not explain why selective ignoring is ineffective. Selective ignoring, by definition, targets the value attributed to the envied items and ignores those values to reduce negative emotions. For example, patients may be trained not to focus on how much they think their classmate's watch is worth. Therefore, if reframing value is the key, then selective ignoring should have similar effects to self-reliance. One response to this contradiction is a methodological argument. Smith and Kim (2007) believe that both self reliance and selective ignoring are "tapping emotional control, perseverance and goal commitment" and are hence effective in coping with envy. They point out that Salovey and Rodin's study is limited to self-reported data, and the assessment of coping strategies should also combine with other measures.

However, Smith and Kim's justification is not strong enough to revive selective ignoring as a suitable remedy for envy. First of all, Salovey and Rodin had only self-reported data. Nevertheless, if selective ignoring and self-reliance were equally effective, the same measurement should have been detected. Second of all, selective ignoring and self-reliance are not identical in their mechanisms. Only self-reliance realizes Aristotle's core idea of *Enkrateia* (self-control); by developing one's own ability without external help, one can directly address the mind state of envy, developing controls over envy rather than suppressing envy. Finally, and most importantly, reframing does not suppress or ignore values through selective ignoring. Reframing allows one to confront old framings and to regain control over the self-regulatory system mentioned before. However, selective ignoring is a way of avoidance by not facing the obstacles that envy creates, which never allows people to examine themselves to overcome malicious envy wisely. Therefore, self-reliance builds up a self-control system

to overcome weakness of will. It enables us to seek better judgments without stagnation or self-destructive behaviour.

Unconscious Difference Monitoring Mechanism Serving an Evolutionary Purpose

In order to better understand malicious envy, it is essential to examine its evolutionary origin. From an evolutionary perspective, malicious envy is based on an unconscious difference monitoring mechanism. Malicious envy is developed to cope with potential threats in order to survive within a social group (Gerhardt, 2009). The specific threat could be the lack of resources or unequally distributed resources for survival, which leads to anxiety. Another threat stems from social comparisons—either being too different compared to others in the group or being at a lower position in the social hierarchy. These two threats can provoke fear of being eliminated through natural selection. Human beings innately developed an unconscious mechanism to constantly monitor the difference between the self and others from the social group in order to evaluate their current position and avoid evolutionary elimination (Gerhardt, 2009).

Depreciating values of resources is an efficient strategy to cope with the inequity of resource distribution. Envy involves ownership of resources, which can be both physical and abstract. The strategy is to depreciate the values of the resource by shifting attention to more salient values. Quintanilla and Jensen de Lopez (2013) studied envy in child development, especially how toddlers experience and resolve envy. They found that mothers from western societies usually use the strategy of distracting their children with a novel object to help them shift attention away from the envied object. This is a specific way of depreciating the original envied object and shifting attention to a novel object, thereby regulating the feeling of envy. It is an intelligent non-stressful way to resolve the feeling of envy without directly suppressing the desire with frustration. Depreciating the value of resources is based on the mechanism of reframing, which reassesses the importance of the envied target. Therefore, although this tendency to envy is embedded in our evolutionary coding, we can prevent the development of malicious envy through educational strategies such as depreciating the value of resources. After all, in developed countries, most residences have sufficient resources for survival, but the fear of a lack of resources still remains. However, evolution also gave us a pow-

erful prefrontal cortex and profound neuroplasticity, which means we are capable of inducing the skill of reframing through self-evaluating and reframing to develop a more adaptive cognitive style.

As said, in unconscious difference monitoring, people are constantly making upward social comparisons. Although such comparisons may lead to envy, they do not necessarily lead to destructive behaviour. Some envy can serve as motivations for self-improvement. Indeed, this mechanism is known as benign envy. Benign envy can stimulate better performance through upward comparisons. Such envy could also be understood as admiration, which provokes a feeling of happiness with an appreciation of how others are good. On the contrary, malicious envy involves unhappiness cultivated by the incapability in improving performance. The incapability of improving performance is equivalent to the idea of inferiority, which results in low self-esteem. Researchers found that malicious envy occurs when two conditions are present: 1) the subject is making an upward social comparison and 2) the subject believes improvement is hard (Ven et al., 2011). Therefore, people with malicious envy cannot motivate themselves to improve performance from upward social comparisons.

Reframing is the essence of depreciation. Depreciating values of resources is a way to reassess the importance that values of resources carry. Furthermore, depreciating values of resources can also resolve low self-esteem. People with envy suffer from low self-esteem due to upward social comparisons, and people develop defensive attitudes to protect their self-esteem from inferiority. Thus, depreciation is a reframing technique to reduce the vital value of the envied target; once the value is decreased, less upward social comparisons will be made, thereby bolstering self-esteem and resolving inferiority. Self-enhancement is developed through this process. It fosters people to become strong in gaining firm will and belief in their capabilities to overcome envy.

Conclusion

The purpose of *reframing* is to defeat self-destructive behaviour and resolve irrational instinct desire. The fundamental coping strategies all derive from reframing based on examination of the mechanisms of malicious envy. Cognitive Behavioural Therapy and mindfulness practice, including meditation, resolve misattribution of salience and inner conflict by relying on insightful reframing to formulate a re-evaluation

of the envied targets. Improving self-reliance allows one to develop a strong will, thereby building up autonomy to reframe the value of envied targets to counter weakness of will. Moreover, by depreciating the value of the envied target, one adjusts unconscious difference monitoring and shifts attention to another salient target to overcome envy. Finally, reframing, as the fundamental core mechanism of the coping strategies, helps people reassess their envied target and irrationality in order to pursue inner peace.

References

Anderson, R. (1987). Envy and Jealousy. *Journal of College Student Psychotherapy, 1*(4), 49-82. doi:10.1300/j035v01n04_04

Baumeister, R. F., Heatherton, T. F., & Tice, D. M. (1994). *Losing control: how and why people fail at self-regulation.* San Diego: Academic Press.

Beck, J. S. (1995). *Cognitive Therapy: Basics and Beyond.* New York: The Guilford Press.

Cotter, P., Meysner, L., & Lee, C. W. (2017). Participant experiences of Eye Movement Desensitisation and Reprocessing vs. Cognitive Behavioural Therapy for grief: Similarities and differences. *European Journal of Psychotraumatology, 8*(Sup6), 1375838. doi:10.1080/20008198.2017.1375838

Davidson, D. (2001). How Is Weakness of the Will Possible? *Essays on Actions and Events,* 21–42. doi:10.1093/0199246270.003.0002

Davidson, K., Norrie, J., Tyrer, P., Gumley, A., Tata, P., Murray, H., & Palmer, S. (2006). The Effectiveness of Cognitive Behavior Therapy for Borderline Personality Disorder: Results from the Borderline Personality Disorder Study of Cognitive Therapy (BOSCOT) Trial. *Journal of Personality Disorders, 20*(5), 450-465. doi:10.1521/pedi.2006.20.5.450

Evans, D. R., Boggero, I. A., & Segerstrom, S. C. (2016). The Nature of Self-Regulatory Fatigue and "Ego Depletion." *Personality and Social Psychology Review, 20*(4), 291–310.doi:10.1177/1088868315597841

Figner, B., Knoch, D., Johnson, E. J., Krosch, A. R., Lisanby, S. H., Fehr, E., & Weber, E. U. (2010). Lateral prefrontal cortex and self-control in intertemporal choice. *Nature Neuroscience, 13*(5), 538–539. doi:10.1038/nn.2516

Frankfurt, H. G. (2005). *On Bullshit*. Princeton: Princeton University Press.

Freud, S., & Strachey, J. (1966). *The complete introductory lectures on psychoanalysis*. New York: Norton Library.

Gerhardt, J. (2009). Reply to Commentaries. *Psychoanalytic Dialogues, 19*(3), 318-335. doi:10.1080/10481880902946146

Gerhardt, J. (2009). The Roots of Envy: The Unaesthetic Experience of the Tantalized/Dispossessed Self. *Psychoanalytic Dialogues, 19*(3), 267-293. doi:10.1080/10481880902946021

Harris, C. R., & Salovey, P. (2008). Reflections on Envy. *Envy*, 335-356. doi:10.1093/acprof:oso/9780195327953.003.0018

Hawkins, D. R. (2001). *The Eye of the I: From Which Nothing is Hidden.* West Sedona, AZ: Veritas Publishing.

Heatherton, T. F., Wyland, C. L., Macrae, C. N., Demos, K. E., Denny, B. T., & Kelley, W. M. (2006). Medial prefrontal activity differentiates self from close others. *Social Cognitive and Affective Neuroscience, 1*(1), 18–25. doi:10.1093/scan/nsl001

Holton, R. (2003). How is Strength of Will Possible? *Weakness of Will and Practical Irrationality*, 39–67. doi:10.1093/0199257361.003.0003

Kalis, A., Mojzisch, A., Schweizer, T. S., & Kaiser, S. (2008). Weakness of will, akrasia, and the neuropsychiatry of decision making: An interdisciplinary perspective. *Cognitive, Affective, & Behavioral Neuroscience, 8*(4), 402–417. doi:10.3758/cabn.8.4.402

Marques, J. (2011). Consciousness at Work: A Review of Some Important Values, Discussed from a Buddhist Perspective. *Journal of Business Ethics, 105*(1), 27-40. doi:10.1007/s10551-011-0932-8

Metzinger, T. (2003). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences, 2*(4), 353–393. doi:10.1023/b:phen.0000007366.42918.eb

Modinos, G., Ormel, J., & Aleman, A. (2009). Activation of Anterior Insula during Self-Reflection. *PLoS ONE, 4*(2). doi:10.1371/journal.pone.0004618

Palumbo, D. (2004). Reframing. *The Lancet, 363*(9419), 1484.doi:10.1016/s01406736(04)16129-1

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making, 10*(6), 549–563.

Perini, I., Gustafsson, P. A., Hamilton, J. P., Kämpe, R., Zetterqvist, M., & Heilig, M. (2018). The salience of self, not social pain, is encoded by dorsal anterior cingulate and insula. *Scientific Reports, 8*(1). doi:10.1038/s41598-018-24658-8

Petchkovsky, L., Petchkovsky, M., Morris, P. W., Dickson, P. D., Montgomery, D. T., Dwyer, J., Burnett, P. E., & Strudwick, M.W. (2011). The fMRI correlates of psychological"complexes". Exploring the neurobiology of internal conflict. *Asia-Pacific Psychiatry. 2*(3). A10-A11.

Pritchard, J. P., Jaeger, W., & Highet, G. (1945). Paideia: The Ideals of Greek Culture. Volume 2. In Search of the Divine Centre. *The Classical Weekly, 38*(11), 86. doi:10.2307/4342045

Quintanilla, L., & López, K. J. (2013). The niche of envy: Conceptualization, coping strategies, and the ontogenesis of envy in cultural psychology. *Culture & Psychology, 19*(1), 76-94. doi:10.1177/1354067x12464980

Rathbone, G. (2012). Envy. British *Journal of Psychiatry, 201*(06), 465. doi:10.1192/bjp.bp.110.089284

Rogers, C. R. (1959). A Theory of Therapy and Personality Change: As Developed in the Client-Centered Framework. *Psychology: A Study of a Science. Study 1, 3,* 184–256.

Salovey, P., & Rodin, J. (1988). Coping with Envy and Jealousy. *Journal of Social and Clinical Psychology, 7*(1), 15-33. doi:10.1521/jscp.1988.7.1.15

Sato, T. (2005). The Internal Conflict Model: A Theoretical Framework for Integration. *The Humanistic Psychologist, 33*(1), 33–44. doi:10.1207/s15473333thp3301_4

Smith, R. H. (1991). Envy and the sense of injustice. In *The psychology of jealousy and envy.*

Smith, R. H., & Kim, S. H. (2007). Comprehending envy. *Psychological Bulletin, 133*(1), 46-64. doi:10.1037/0033-2909.133.1.46

Smith, R. H., Combs, D. J., & Thielke, S. M. (2008). Envy and the Challenges to Good Health. *Envy,* 290-314. doi:10.1093/acprof:oso/9780195327953.003.0016

Tabak, N. T., Horan, W. P., & Green, M. F. (2015). Mindfulness in schizophrenia: Associations with self-reported motivation, emotion regulation, dysfunctional attitudes, and negative symptoms. *Schizophrenia Research, 168*(1-2), 537-542. doi:10.1016/j.schres.2015.07.030

Ven, N. V., Zeelenberg, M., & Pieters, R. (2009). Leveling up and down: The experiences of benign and malicious envy. *Emotion, 9*(3), 419-429. doi:10.1037/a0015669

Ven, N. V., Zeelenberg, M., & Pieters, R. (2011). Why Envy Outperforms Admiration. *Personality and Social Psychology Bulletin, 37*(6), 784-795. doi:10.1177/0146167211400421

Vervaeke, J., Lillicrap, T. P., & Richards, B. A. (2012). Relevance Realization and the Emerging Framework in Cognitive Science. *Journal of Logic and Computation, 22*(1), 79–99. doi:10.1093/logcom/exp067

Wheeler, L. (1966). Motivation as a determinant of upward comparison. *Journal of Experimental Social Psychology*. https://doi.org/10.1016/0022-1031(66)90062-X

Young, R. (1987). Egalitarianism and envy. *Philosophical Studies, 52*(2), 261-276. doi:10.1007/bf00646459

Zeng, X., Chan, V. Y., Liu, X., Oei, T. P., & Leung, F. Y. (2017). The Four Immeasurable Meditations: Differential Effects of Appreciative Joy and Compassion Meditations on Emotions. *Mindfulness, 8*(4), 949-959. doi:10.1007/s12671-016-0671-0

I chose to fly.

My feet replanted to the ground.
And psyche won't let go.
A bleeding heart.
Drips of hidden
tears.
In need of...

Freedom.

You said to me once

There's a
darkness
that lives
inside me.

I said to you once

I hope that you can remember
love in your darkest moments
and that it will transcend time
and distance and lift you up.

But right now.
We're just treading water for our lives.

If one is falling.
The cave will never swallow.
But.
It will not let go.

We came. We went.
Valentyn Korotkevych and Renée Mak

A Morality Module for Machines

Emily J. Davidson
York University

This paper will seek to explore some of perceived barriers to the development of morality in digital, algorithmic form. These impediments tend to fall within two categories: practical problems and ethical or moral issues. First, I will articulate the prevailing positions on these difficulties and subsequently express why one of the established proposals, the evolutionary paradigm, ought to be considered the optimal substratum upon which to build a moral architecture in light of such issues. This is best conceptualized through the development of an independent moral algorithm that may be implemented in a diverse range of automata, effectively converting these machines into artificial moral agents. New innovations in deep learning may facilitate such a project in ways that have not previously been possible so as to belay many of the practical problems raised as objections to artificial, moral decision-making. I will outline a rudimentary springboard for the development of such an algorithm and subsequently discuss how, generated in the ways I will describe, this algorithm may tackle some of the preeminent roadblocks pertinent to the development of moral algorithms.

Introduction

History is littered with convictions about what science and technology can never hope to achieve, but retrospect tends to judge such statements unfavorably. In 1842, Auguste Comte, considered by many to be the father of modern scientific philosophy, made the claim that we would *never* know the chemical composition of stars (Comte, 1842). It is easy to look upon such an oversight and sympathize with an inability to predict astronautics. Spectroscopy (the means through which the chemical compositions of stars were first identified), however, had

in fact been long established in Comte's time; it simply had not yet proven its *astronomical* potential. Comte's blunder demonstrates that even our wildest assertions about what can be accomplished often hinges upon creativity and innovation in using technologies already available. The question of artificial intelligence and morality shares a similar story. Dialogues certainly exist on the costs and benefits of moral machines, but much of the focus fixates on *whether it is even possible* to algorithmize morality at all. I wish to suggest that such a distraction detracts from the very necessary conversations that must be had as we stand on the precipice before artificial intelligence that may soon surpass our own. Nearly all of the philosophical questions on artificial intelligence from super-soldiers to bioethics have shifted the dialogue to '*when* it emerges' long ago. I wish to suggest that the same movement ought to occur with moral machines. New techniques in deep learning have made radical shifts in the way computers process information, making them more like human brains than ever before. I wish to make the case that moral machines are not only imminent, but possible with the tools that we have now. I wish to explore the possibility in the later half of this work, by outlining a crude design of what an algorithm for morality might look like in a machine. It is perhaps these new programming tools that may inspire the next generation of developers to bypass the Comte trap and unlock a new world of machines capable of building values just like our own.

Exploring Our Current Position in the Artificial Moral Landscape

The Frontier of Artificial Morality: If we assume that the algorithmizing of morality is possible, I wish to suggest that would be ethically *irresponsible* not to do so, for two major reasons. First, if we assume that, as artificial intelligence and machine proficiency progresses, it will begin taking over many of the roles and responsibilities currently allotted to humans, then we can assume that these machines will be tasked with decision-making that can affect our well-being or survival. Implementation of such agents in the absence of any moral or ethical framework is bound to lead to unforeseen consequences of harm, whether physical or psychological. This has become a mainstream issue with the rollout of self-driving cars. Unfortunately, much of the discussion about self-driving cars has been heavily focused on the "hard" cases: would a person prefer a car that would sacrifice a driver

or pedestrian should such a scenario arise? Studies demonstrate that on these types of cases, there is likely to be little universal consensus (Awad, et al., 2018). Such an emphasis on these dilemmas make for interesting philosophical discussion, but are on the whole, steering attention in the wrong direction. The likelihood of humans entering such scenarios of *unavoidable* fatal collision are *extremely* remote; the overwhelming majority of the ethical decisions made behind the wheel involve proactive prevention of avoidable collisions (National Highway Traffic Safety Administration, 2020; The Brown Firm, 2018; World Health Organization, 2020). This boils down to trivialities like choosing to obey speed limits or stop signs on sparsely populated roads, the kind of mindless decisions machines excel at.

Second, domain-specific machines simply have more computational resources dedicated to a single facet of knowledge, thus lending them to significantly high degrees of accuracy in their domain. Humans, as marvelous domain-general computation machines, sacrifice accuracy for versatility. Therefore, in domain-specific circumstances, we can expect the machines designed for these purposes to be more accurate overall, leading to a responsibility in miitigating harm resulting from human error. This can be more difficult to digest as it requires not only the assumption that artificial morality is possible, but that this moral framework could be *superior* to that of humans. Even if one remained skeptical of superiority in *theoretical moral capacity*, it is difficult not to acknowledge the ways in which human error might prove to be a crippling confound making even theoretical moral superiority moot. Artificially intelligent vehicles eliminate the variables of fatigue, distraction, intoxication, and health complications while driving, just to name a few, and the removal of the decisions these variables engender will likely result in more ethical decision-making on the whole (Cuneen, Mullins, & Murphy, 2019). Moreover, the likelihood of a human carefully selecting and enacting the optimal course of action faster and more accurately in unavoidable collision scenarios than a domain-specific, intelligent driving system invites skepticism. Faster processing and smaller margins of error are simply more likely to be successful mitigating harm in these scenarios.

On the road towards technological advancement, machines will continue making more decisions with increasing importance to our well-being. Thus, any reduction in the negative impacts upon well-being overall is worth considering as an incentive for instantiating moral

capabilities in artificial intelligence, irrespective of developmental sophistication. If such an algorithm sufficiently reduces overall harm by even a fraction of a percent, then this is still, almost by definition, a net-benefit. Therefore, provided that such tools exist, they ought to be instantiated *immediately*. As such, the remainder of this paper shall discuss the current barriers for implementation and some of the ways in which they might be overcome.

Current Roadblocks to Algorithmizing Morality: Suppose we avoid the Comte trap and agree that it is theoretically possible to implement artificial morality and further, that such an algorithm can be generated with technology already available to us. This tells us little about the specifics of how to go about actually developing and instantiating such an algorithm, not to mention the difficulty of legitimization. These kinds of practical problems tend to align with four major issues. First, and perhaps most obvious, is the lack of a universally accepted set of parameters by which to define 'morality'. The subjectivity of the material makes the coding of that which cannot be codified problematic (Allen, Varner, & Zinser, 2000). Humans are, after all, the ultimate moral arbiters; if they do not recognize the algorithm as acting under moral principles, then this completely undermines its legitimacy as a moral agent.

The second difficulty is the enticement towards the utilization of the moral framework when making decisions and acting within the world. What incentivizes an agent to abide by moral rules when such rules run in contravention to its goal states (Allen, Varner, & Zinser, 2000)? In humans, this can often be facilitated through normative social expectancies. Machines, however, lack any biological imperative towards social affiliation. The difficulty of enticement in the absence of a socioemotional toolkit converges onto the third problem: the degree to which emotions are necessary for morality. Many philosophers and neuroscientists alike speculate that emotions, despite their potential in generating immoral action, might still be necessary for moral action of any kind (Damasio, 2004; Coeckelbergh, 2010; Prinz, 2011). Emotions facilitate *independent decisions* by generating valence heuristics for action towards an ultimate purpose, enabling action in completely novel scenarios (Kavathatzopoulos & Asai, 2013). An autonomous agent without a valence heuristic may be unable to interpolate the meaningful distinctions required to autonomously evaluate unforeseen ethical scenarios.

The fourth and final issue concerns the role of empathy. The degree to which it is necessary for morality is hotly debated, but the arguments against it tend to lean on idealized characterizations of empathy, rather than practical ones. Even the fiercest advocates of removing empathy from the moral equation acknowledge that humans do, in practice, use empathy to make moral evaluations (Prinz, 2011). It remains unclear, however, how even idealized moral systems without empathy could be capable of executing any of the major moral systems we recognize. Moral systems *without* empathy may one day be possible, but at present, the human execution of morality involves the use of empathy, and it is unclear what a moral system without this ability would entail. Thus, if for no other purpose than universal acceptance as a moral agent, artificial moral intelligence will likely be required to display some form of empathy.

Evolution's Solution: It seems unlikely that a universal, moral system formulated as a series of *a priori* rules and guidelines can readily be generated for two major reasons. First, there is simply no consensus around prescriptions for moral behaviour. Second, a moral system must be able to accommodate novel scenarios for which no explicit rule has been established. The fact that humans are capable of acting morally in novel scenarios, even in the absence of explicit frameworks, suggests that a top-down codification of moral rules may be unnecessary for creating moral machines, and, perhaps, even understanding morality itself. Our moral faculties, like most human attributes, are products of bottom-up evolutionary forces, billions of years in the making. While we may not all share moral tenets, each of us shares the evolutionary trajectory of our moral hardware. Therefore, the generation of an algorithm with all of the autonomy, flexibility and variability present in human morality is likely best served through a bottom-up approach that is reinforced based on environmental feedback.

In order to make use of a bottom-up approach to morality, the algorithm will require some kind of explicit criteria towards which it aims, such that a cost-benefit analysis might be conducted for decision-making. Like most bottom up systems, this will hinge on success or failure as contextualized by its environment. The difficulty, however, is determining what *constitutes* success or failure. Resolving this issue forces the imposition of a top-down success criterion for moral behaviour that may not be universally accepted. I wish to claim, how-

ever, that there is at least one necessary condition converged upon by all moral paradigms: they all appear to involve the interaction with an expectation between, at minimum, two participants. Participants can involve people, animals, deities or even, in many cases, intangible constructs, as long as an expectation can be shared in some form between them. The interaction of the participants within this expectation can ultimately result in some form of moral evaluation. This condition, as it happens, also heavily resembles what is typically defined as 'cooperation'. Therefore, in an effort to better understand how to develop morality from the bottom-up, it makes sense to begin with the evolution of cooperation.

In 1980, Axelrod solicited strategies to be pitted against each other in a tournament based on an iterated game called the Prisoner's Dilemma. The Prisoner's Dilemma paradigm is one in which there are two 'players,' and each has the ability to either 'cooperate' or 'defect' on each round. The overall cost or benefit to a player will be the direct result of two factors: the decision made by the first player and the decision made by the second player. Two versions of this tournament were held, the first of which was limited to 200 rounds. The second, in an effort to better match the uncertainties in organic cooperative behaviour, left the number of rounds unknown (Axelrod & Hamilton, 1981). In both versions of this tournament, a strategy submitted by Anatol Rapoport emerged victorious: TIT FOR TAT (Axelrod & Hamilton, 1981). His strategy was fairly simple: it opened with a cooperative move, and then directly matched the move made by its opponent in the subsequent round.

Axelrod, along with evolutionary biologist William D. Hamilton, published these findings in a landmark paper, *The Evolution of Cooperation*, which discusses the reasons TIT FOR TAT was so successful. TIT FOR TAT's success was surprising, given that it began with a cooperative move, thus leaving it open to exploitation by defectors. Mathematically, however, it was more successful over the long term, provided that the possibility of mutual cooperation between itself and another strategy existed (Axelrod & Hamilton, 1981). This led to the revelation that *reciprocal* altruism could lead to superior evolutionary outcomes than purely selfish behaviour. Operating under such a strategy would provide organisms a disproportionate advantage in meeting survival and reproductive needs, provided that these organisms had a

high likelihood of repeated encounters and some mechanism by which to individualize and punish defectors when reencountered. The better an organism fit such criteria, the more readily cooperative, stable strategies arise (Axelrod & Hamilton, 1981). Additionally, the more complex societies become, the more tangible rules for cooperation and defection behaviour become, as they begin to coalesce into moral, and subsequently legal, systems (Curry, 2016).

Reciprocal altruism as described by Axelrod is not the only theory on the evolution of cooperation. It is, however, the most widely accepted, and more importantly for the purposes of AI development, it works. Bottom-up computer models can be generated with cooperation strategies emerging, both in the Prisoner's Dilemma and in other games (Ale, Brown, & Sullivan, 2013). Thus, for the purposes of generating a bottom-up, machine learning algorithm for morality, reciprocal altruism appears to be the most appropriate foundational success criterion. Simple teaching scenarios, converted into game theory, can be presented to the algorithm while more complex mechanisms, like the coding of scenarios in human language, can be introduced over time. The training would seek to first mimic cooperation as it evolved in organisms from the simple to the complex, and subsequently, use this foundation to learn morality in the same way a child. Eventually, much like the occurrence in humans, the machine learning process will begin to make inferences across scenarios that will act as overarching principles to guide decision-making in novel cases once the training phase is complete.

The development of a bottom-up, moral framework in similar fashion to human moral learning helps deal with the first roadblock in the development of an independent moral algorithm. If the machine is generating the ability to be moral from the ground up, it is not necessary to worry about defining and codifying moral rules as a precursor to development. The machine will recapitulate the evolutionary moral trajectory from simple organisms to human infants, and subsequent moral training from human infants to ethical adult agents. While this process took billions of years in humans, in machines learning, virtually all of the latency time associated with experiential learning is eliminated. The algorithm can be exposed to millions of game iterations in just days, and this number only increases with the advancement of technology. Training will use a tree sampling search parameter similar

to that of AlphaZero in order to modify the learning target with each improved state. This allows for cost/benefit targets to be continually increased as training opportunities demonstrate them to be possible (Carlsson & Öhman, 2019).

Developing a Moral Algorithm with Machine Learning Techniques

The algorithm described here is meant to be supplementary, rather than substantive, meaning that the goal is to incorporate it into existing machine software, rather than develop a functioning machine with this algorithm alone. Further, upon subsumption, a balance must be struck between integration with the features of its host and the maintenance of domain specificity in cooperation calculations. The integration of the host goal-states must be incorporated without eclipsing the cooperation calculations, calculations which should ideally supersede host goal-states if the algorithm is to be effective. There will need to be a way in which to make moral decision-making sufficiently salient to a machine so as to override existing goal-states, in similar fashion to the function of a human conscience.

Digitizing the Conscience: The discussion of a digitizing psychological phenomena like a conscience may seem strange, but the collaboration between the brain and computer sciences have led to some major breakthroughs in artificial intelligence. One of the most exciting and effective solutions emergent from such partnerships is reinforcement learning, a method of machine learning that directly mimics the way human brains learn via the dopaminergic system (Sutton & Barto, 2018). The dopaminergic system contributes to learning through evaluation of potential reward versus actual reward, and then reconfigures synaptic weights accordingly so as to find a more accurate balance between the two variables (Glimchar, 2011). This is explicitly what the digital form of reinforcement learning does. But the dopaminergic system is more than a reconciliatory mechanism for reward prediction; it is also generative. 'Reward', as measured by the network, can mean either *potential* or *actual* reward (or both), depending simply on which nodes are active. This allows it to be a semi-closed system for reward-based learning. This reinforcement effect is bidirectional and multiplicative in that both positively and negatively valanced neurochemical triggers can be generated. This is based not only on actual reward but also the discrepancy between actual reward and predicted

reward, with increased feedback-related negative affect as actual reward falls below the predicted (Bismark, Hajcak, Whitworth, & Allen, 2012; Schulz, Dayan, & Montague, 1997). Put simply, reinforcement learning both biological and digital do the following:

1. Assign a positive value to high reward and negative value to low reward
2. Generate reward by reducing discrepancy between current state and desired state
3. Generate reward-value based predictions for available decision options
4. Calculate the difference between predicted reward and actual reward
5. Solve for the appropriate reward value and reconfigure connective weighting accordingly

Perhaps the most important part of this picture is that reinforcement learning captures the current landscape of neural connectivity, whether biological or digital and in outcomes of reward (whether positively or negatively sourced), strengthens (or weakens, for negative reward) connectivity to *all* connections that are active at the time of reinforcement (Sutton & Barto, 2018). This entails that many individual nodes or neurons will be inappropriately reweighted in each instance of reinforcement and thus, the more reinforcement trials calculated, the more accurate the system will become. This is where machines hold an advantage over the human system. In humans, action times, energy constraints, and reward resources radically limit the number of targeted reward trials that can be tested in any given timeframe. For example, in order for a human to *accurately* learn whether cheating on a test is more likely overall to yield negative rewards, many instances of trial and error must occur to lessen the influence of idiosyncratic variables and results. The ability for a human to complete many trials of cheating on a test is severely limited for obvious reasons. An algorithm in training, even on suboptimal hardware, can run through over a million observations over the course of a day.[1] This suggests that moral training can occur extremely quickly in a machine.

1 This figure is based upon conservative estimates from the observations per day achieved by machines like AlphaZero in addition to observation numbers per hour provided by novice and expert coders who have generated machine learning algorithms. The number of variables in the training set will act as the main determinant of training time with standardized computing power.

Reinforcement outcomes in a machine must serve a greater purpose than merely expediting training times and minimizing error rates. If the reinforcement mechanism does not sufficiently impel the machine to make ethical decisions, it fails to serve as a moral framework because competing goal-states will inevitably take precedence in machines originally developed to accomplish certain aims. The question that must be resolved is how to make cooperation highly rewarding. Neuroscience research on this very topic has demonstrated that perhaps this is already the case. Cooperative behaviour does, in fact, trigger dopaminergic reward (Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004; Decety, Jackson, Sommerville, Chaminade, & Meltzoff, 2004); more specifically, cooperative behaviour in game theoretical parameters of the Prisoner's Dilemma (Rilling, et al., 2002). If our dopaminergic reward system has been forged by evolution to facilitate goal states of survival and genetic reproduction, why would choosing cooperation over defection trigger such a surge in dopaminergic activity? This harks back to Axelrod and Hamilton's findings: even goal-states that are 'selfish' in nature can be more readily attained through cooperative means.[2] Thus, while the output of the machine's training period is simply a raw reward signal, the data sets for training implicitly teach the machine that cooperation is rewarding *because* it facilitates maximizing its own goal-states. Thus, when implemented into a machine designed for potentially conflicting goal-states, the algorithm offers the learned principle that cooperation will better facilitate individual goal-state satisfaction overall. Moreover, based on the evidence provided by Axelrod and Hamilton that this is indeed the case, the choice of cooperation in an effort to secure selfish goal-states will continue to be reinforced, provided that the machine operates within environments in which at least some cooperative agents are present. As a result, the module will be able to act as a moral guide based on learned principles much like a human conscience may learn from intrinsic evaluations of real and potential actions (Pitrat, 2009).

Emotionality Sans Endocrinology: There is a school of thought in philosophy that supposes emotions to be a crippling force in making moral decisions, holding a Spock-like creature as a paragon of moral arbitration. This school of thought is typically populated by utilitar-

2 The term "selfish" in this context is meant to convey the procurement of individual benefit with respect to individual goal-states. It is not mean to convey any sense of moral intentionality.

ians, and indeed, there is evidence to suggest that such thinkers are less emotionally driven in moral decision-making. Links have been found between philosophical perspectives on morality and activation of anticorrelated regions involving reasoning and emotional response. For example, people framing the trolley problem as utilitarians show greater activation of the 'rational, cognitive' dorsolateral prefrontal cortex (dlPFC) (Greene, Somerville, Nystrom, Darley, & Cohen, 2001). It is unclear, however, whether emotions can be eliminated from moral calculations overall. As the difficulty of the moral dilemmas increases, so too does the incidence of dlPFC activation give way to the emotional centers, and incidences of utilitarian decision-making drop off (Greene, Nystrom, Engell, Darley, & Cohen, 2004). Learning from past moral calculations can lead to semanticized principles that do not heavily draw on emotional states, but this does not mean those emotional states were not necessary for creating them (Haidt, 2001; Prinz, 2007). Moreover, modifying emotional elements of affectivity, empathy, and motivational impetus to act change moral evaluations of action (Manfrinati, Lotto, Sarlo, Palomba, & Rumiati, 2011; Tangney, Stuewig, & Mashek, 2007; Ugazio, Lamm, & Singer, 2012). Emotions appear to be an integral part of the moral evaluation puzzle and it is unclear how they might be excised while still maintaining the integrity of the term 'morality'.

Even if moral systems without emotions could be empirically demonstrated, an algorithm for mainstream acceptance and use will require decision and evaluation criteria recognizable to the average user. For example, the non-consensual injection of disease or oncogenic cells into healthy adults for the purpose of testing new vaccines and treatments could likely speed up the breakthroughs of medical advances, benefitting wide swathes of the population. The average person understands, however, that the deaths resulting from such a practice would be an unacceptable transaction. An artificial moral agent advocating these steps on a purely utilitarian platform would likely be viewed by the overwhelming majority of the population as amoral.[3] Even extreme utilitarianism must recognize that the overall reduction to human wellbeing elicited by the suffering of others must be a factor in the calculation of *total net* wellbeing. Most forms of util-

3 Evidence of such a claim can be viewed in the literature with respect to psychopathy. 'Cold', calculating decision-making with implications on moral outcomes tend to be classified as 'amoral' rather than 'immoral'.

itarianism, however, have difficulty dealing with *subjective* suffering, both on theoretical *and* practical grounds. The more difficult the subjective suffering calculation for wellbeing becomes, the more we must rely on bottom-up processes to attempt a reconstruction of subjective suffering, such that we might then try to calculate it. This may be why some regions of the brain that contribute to rule encoding (dlPFC) are anticorrelated with those more involved in valence evaluation, such as the ventromedial prefrontal cortex (vmPFC) (Mian, et al., 2014; Nicole & Goel, 2014). The vmPFC attempts a bottom-up reconstruction of emotional salience. These calculations are then offloaded to the dlPFC to create principles (a beneficial shortcut to save on metabolic costs). However, when no principle reconciles our 'gut-feelings', we tend to revert back to bottom-up emotional decision-making that involves emotional salience calculations (Nicole & Goel, 2014). The method by which evaluations and decisions are reached digitally will likely have to be similarly calculated in order to gain initial, widespread acceptance.

Even if it were possible to generate moral systems that were accepted as such by the general population *without* any kind of surrogate for emotions, there is yet another reason the inclusion of one may be desirable: efficiency. The sheer number of variables and complexities that factor into meaningful decision-making is astronomically high. Emotional impulses were the gold standard of cognition through much of our evolutionary trajectory because they were a fast and metabolically cheap way of coordinating action. In this way, emotions can be viewed as specializations of cost function calculation, whereby costs and benefits can be determined quickly and efficiently at the expense of accuracy when such an approach would be beneficial (Marblestone, Wayne, & Kording, 2016). This is likely a necessary feature if the module is expected to make evaluations quickly and in real-time alongside other software. If a mechanism is able to fulfill this role in a similar fashion, it may be sufficient to function as a surrogate for emotion in digital space.

New deep learning techniques called 'autoencoders' may be capable of accomplishing this goal. The basic idea of a standard autoencoder is that it takes a set of input nodes and tries to convert the configuration into a sufficiently low-resolution version of the information. It seeks to filter out information from the inputs through an 'encoder' to generate increasingly low-resolution versions so that this information might

pass through a bottleneck consisting of a significantly reduced number of nodes from the original input set. From the compression state of the bottleneck, the information passes through a 'decoder' which aims to recreate the information lost in the encoder phase and submit it to an output. The better the output expresses the input, the more effective the autoencoder. The benefit of this process is that it allows for the generation of compressed versions of a data set that capture only the most salient information necessary for use. This mechanism, in fact, looks quite similar to what neurotransmitter systems are doing in the nervous system.

Glucocorticoid hormones involved in the stress response, offer a relatively simple illustration of the analogous role in biological organisms. Suppose a person is walking through the forest and hears a loud growling coming from directly behind them. Virtually all of the input information being received in that moment become impediments to expedient decision-making. Were this individual to spend precious seconds noticing the colour contrast in the leaves or the coolness of the breeze, they likely would not have lived to propagate as many offspring. Evolution has selected for the ability to filter out stimuli from a large set of inputs and compress the data to the most relevant, salient features. A low-resolution version of the inputs, cortisol, can then move quickly and effectively throughout the body to convey the pertinent message of 'danger' and upon decoding, contextual information provides cues as to how movement ought to occur and what steps are to be taken. These elements are encompassed in the analogous components of an autoencoder, complete with input system (activated neural network), encoders (pre-synaptic neuron locations), bottleneck (molecules released), decoder (post-synaptic neuron locations) and output set. Indeed, newer studies show that this configuration is more plausible than fixed categories of emotion (Skerry & Saxe, 2015; Dubois & Adolphs, 2015), making bottom-up learning of emotional heuristics appear to be closely akin to the mechanisms used by the brain (Marblestone, Wayne, & Kording, 2016).

Uploading Empathy: Empathic machines often strike people as wildly unintuitive. After all, machines do not have biological substrata, so how could they ever *truly* embody human states and understand our perspective? While this might be true, it is equally true to say that no *human* could ever truly embody the state of another human, because

for all their biological similarities, experience of events will likely differ significantly. If, however, empathy is the ability to vicariously experience the emotional state of another, machines may end up superior to humans in this respect, provided that an appropriate surrogate for emotional experience (like the one listed above) is present. This boils down to the different methods of learning in machines versus humans.

Empathy is often framed as pure, adulterated emotional response when in reality, there are both cognitive and emotional components that are equally necessary. In fact, when the pain is emotional rather than physiological, there is more activation of the aforementioned reasoning center, the dlPFC (Sapolsky, 2017). In brains, stimuli regularly feed information in bottom-up format towards emotional centers. When empathizing, however, a top-down reconstruction of abstract variables must be made in order to piece together the emotional picture. This results in a great deal of signal loss when attempting to replicate the emotions of others due to both the top-down nature and heightened cognitive load required. This problem is compounded when considering the anticorrelation between the rational dlPFC and emotionally pertinent vmPFC. Emotional activation must therefore be weakened in order to spare the cognitive resources for interpreting contextual cues (Sapolsky, 2017). As humans practice empathy, they inevitable experience degrees of emotional signal loss.

Unsupervised machine learning, on the other hand, trains algorithm against itself; in Prisoner's Dilemma scenarios, for example, it is both player 1 and player 2. Therefore, because it is not required to waste computing power deciphering the discrepancy in self-other contextual information, its resources are readily available for use in emotional networks and learning. Emotional salience remains the same for both self and other because, during training, these entities are one in the same. As such, it would be trained on millions of game iterations, each time learning that its opponent's emotional status is equally valuable to its own, a principle of perfect empathy. When self-training is effectively complete and training with real individuals begins, this principle still undergirds its knowledge base. The weight given to the values and emotional states of others, could, for this reason, be stronger than that given by the average human, with a caveat. If the machine is meant to account for subjective differences between its own experience of a scenario, and that of another, it too, like humans, will be forced to

dedicate more resources to the interpretation of contextual cues. This, however, may pose significantly less difficulty for machines for two reasons. First, and most obviously, artificial programs can compute domain-specific information significantly faster than humans can, a skill that increases with each passing year. Second, the fact that the operations of the relevant sectors in the prefrontal cortex are anticorrelated in humans may not need to be true of machines. It is entirely possible that these functions may end up with a positively correlated relationship when unconstrained by neuroanatomical or metabolic limitations. In any case, it is likely impossible to expect *perfect* empathy, even from machines, but it is within the realm of possibility that they may simply end up better at it than humans are.

Conclusion

The model outlined above, while rudimentary, might suggest fertile ground in the artificial instantiation of moral and ethical frameworks for intelligent machines. Such a paradigm is meant to offer solutions available to us right now that could stand to improve the overall wellbeing of humans. The way in which such an algorithm would be instantiated and used in the plethora of possible machines may be beyond the scope of this work, but this cannot be discovered without beginning the process of trial and error. The sooner such a project begins realization, the sooner the training processes can begin expanding and the more effective such algorithms can become. Moreover, beginning the training process now can help prevent humans from being blindsided by the inevitable discovery of intelligent machines that have been programmed in the absence of a moral framework. Inoculation against such an outcome may prevent what could otherwise be a disaster for both our species and our planet.

References

Ale, S. B., Brown, J. S., & Sullivan, A. T. (2013). Evolution of Cooperation: Combining Kin Selection and Reciprocal Altruism into Matrix Games with Social Dilemmas. *PLOS One*. doi:https://doi.org/10.1371/journal.pone.0063761

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence, 12*(3), 251-261.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., . . . Rahwan, I. (2018). The Moral Machine Experiment. *Nature*, 59-64.

Axelrod, R., & Hamilton, W. D. (1981). The Evolution of Cooperation. *Science, 211*(4489), 1390-1396.

Bismark, A. W., Hajcak, G., Whitworth, N. M., & Allen, J. J. (2012). The role of outcome expectations in the generation of the feedback-related negativity. *Psychophysiology, 50*(2), 125-133.

Carlsson, F., & Öhman, J. (2019). AlphaZero to Alpha Hero: A pre-study on Additional Tree Sampling within Self-Play Reinforcement Learning. *KTH, School of Electrical Engineering and Computer Science.*

Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology, 12*, 235–241.

Comte, A. (1842). *The Positive Philosphy.* (H. Martineau, Trans.) Kitchener.

Cuneen, M., Mullins, M., & Murphy, F. (2019). Autonomous Vehicles and Embedded Artifical Intelligence: The Challenges of Framing Machine Driving Decisions. *Applied Artifical Intelligence, 33*(8), 706-731.

Curry, O. S. (2016). Morality as Cooperation: A Problem Centered Approach. In T. K. Shackleford, & R. D. Hansen, *The Evolution of Morality* (pp. 22-51). Springer.

Damasio, A. (2004). *Decartes' Error: Emotion, Reason, and the Human Brain.* New York: Penguin House.

Decety, J., Jackson, P. L., Sommerville, J. A., Chaminade, T., & Meltzoff, A. N. (2004). The neural bases of cooperation and competition: an fMRI investigation. *Neuroimage, 23,* 744-751.

Dubois, J., & Adolphs, R. (2015). Neuropsychology: How Many Emotions Are There? *Current Biology, 25*(15), 669-672.

Glimchar, P. W. (2011). Understanding dopamine reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings in the National Academy of Sciences*, 15647-15654.

Greene, J. D., Somerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgement. *Science, 2105*(293), 2105-2108.

Greene, J., Nystrom, L., Engell, A., Darley, J., & Cohen, J. (2004). The neural bases of cognitive conflict and control in moral judgement. *Neuron, 44*(2), 389-400.

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgement. *Psychological Review*(108), 814-834.

Kavathatzopoulos, I., & Asai, R. (2013). Can Machines Make Ethical Decisions? *IFIP International Conference on Artificial Intelligence Applications and Innovations* (pp. 693-699). Berlin: Springer.

Manfrinati, A., Lotto, L., Sarlo, M., Palomba, D., & Rumiati, R. (2011). Moral dilemmas and moral principles: When emotion and cognition unite. *Cognition and Emotion, 27,* 1276-1291.

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an Integration of Deep Learning and Neuroscience. *Frontiers in Computational Neuroscience,* 94.

Mian, M. K., Sheth, S. A., Patel, S. R., Spiliopoulos, K., Eskandar, E. N., & Williams, Z. M. (2014). Encoding of Rules by Neurons in the Human Dorsolateral Prefrontal Cortex. *Cerebral Cortex, 24,* 807–816.

National Highway Traffic Safety Administration. (2020). *www.nhtsa.gov*. Retrieved 2020, from https://www.nhtsa.gov/risky-driving

Nicole, A., & Goel, V. (2014). What is the role of the ventromedial prefrontal cortex in emotional influences on reason? In *Emotion and Reasoning* (pp. 154-173). London: Psychology Press.

Pitrat, J. (2009). *Artificial Beings: The Conscience of a Conscious Machine.* Hoboken, New Jersey: Wiley.

Prinz, J. (2007). *The Emotional Construction of Morals.* Oxford: Oxford University Press.

Prinz, J. (2011). Is Empathy Necessary for Morality? In *Empathy: Philosophical and Psycholoigcal Perspectives* (p. 213). Oxford: Oxford University Press.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A Neural Basis for Social Cooperation. *Neuron, 35*(2), 395-405.

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *NeuroReport, 15*, 2239-2243.

Sapolsky, R. M. (2017). *Behave: The Biology of Humans at Our Best and Worst.* New York: Penguin Books.

Schulz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science, 275*(5306), 1593-1599.

Skerry, A., & Saxe, R. (2015). Neural representations of emotion are organized around abstract event features. *Current Biology, 25*(15), 1945-1954.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (Second ed.). Cambridge, Massachusetts: The MIT Press.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral Emotions and Moral Behaviour. *Annual Review of Psychology, 58,* 345-372.

The Brown Firm. (2018, December 19). *JDSupra*. Retrieved from JDSupra: https://www.jdsupra.com/legalnews/leading-causes-of-car-accidents-with-60370/

Ugazio, G., Lamm, C., & Singer, T. (2012). The role of emotions for moral judgments depends on the type of emotion and moral scenario. *Emotion,* 579–590.

Wallach, W., & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong.* New York: Oxford University Press.

World Health Organization. (2020, February). *www.who.int*. Retrieved March 12, 2020, from World Health Organization: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries