Conceptual control:
On the feasibility of conceptual engineering

Eugen Fischer
University of East Anglia

This paper empirically raises and examines the question of 'conceptual control': To what extent are competent thinkers able to reason properly with new senses of words? This question is crucial for conceptual engineering. This prominently discussed philosophical project seeks to improve our representational devices to help us reason better. It frequently involves giving new senses to familiar words, through normative explanations. Such efforts enhance, rather than reduce, our ability to reason properly, only if competent language users are able to abide by the relevant explanations, in language comprehension and verbal reasoning. This paper examines to what extent we have such 'conceptual control' in reasoning with new senses. The paper draws on psycholinguistic findings about polysemy processing to render this question empirically tractable and builds on recent findings from experimental philosophy to address it. The paper identifies a philosophically important gap in thinkers' control over the key process of stereotypical enrichment and discusses how conceptual engineers can use empirical methods to work around this gap in conceptual control. The paper thus empirically demonstrates the urgency of the question of conceptual control and explains how experimental philosophy can empirically address the question, to render conceptual engineering feasible as an ameliorative enterprise.

Conceptual engineering, experimental philosophy, polysemy, stereotypes, philosophical methods.

1. Introduction

> 'When *I* use a word,' Humpty Dumpty said, in rather a scornful tone, 'it means just what I choose it to mean — neither more nor less.'
> 'The question is,' said Alice, 'whether you *can* make words mean so many different things.'
> 'The question is,' said Humpty Dumpty, 'which is to be master —that's all.' (Lewis Carroll, *Through the Looking Glass*)

This paper will address Alice's question, as it arises for conceptual engineering. The question is one of the most fundamental questions facing this philosophical project, with consequences for the project's feasibility and approach. This paper will contribute to rendering the question empirically tractable, by drawing on psycholinguistics, and will begin to develop an empirically supported answer to it, by drawing on recent work from experimental philosophy. This experimental work lets us get clearer on the extent to which we are masters – rather than at the mercy – of our representational devices. The findings have concrete methodological consequences for conceptual engineering.

Conceptual engineering is concerned with the assessment and improvement of concepts and other representational devices (Cappelen 2018). It seeks to create or adapt such devices for new purposes or to identify and repair defects in them (op. cit.). Relevant defects include cognitive defects that undermine our ability to reason properly, moral or political defects that undermine our ability to act in line with our moral or political values, and semantic defects where terms are empty or engender paradoxes (Cappelen 2018, pp.9-38). For centuries, scientists and philosophers have introduced new concepts (e.g., *gene*, *entropy*, *schizophrenia, human capital*) or new definitions (like Newton's second law as definition of 'force', or truth-table definitions of logical connectives), for a wide variety of purposes, in and beyond research. The first generally acknowledged paradigm of conceptual engineering as an explicit philosophical research program is, however, provided only by Rudolf Carnap's program of formal explication. Carnap used formal methods to remedy the cognitive defects of vagueness and ambiguity for terms relevant for the natural or formal sciences, in order to increase the terms' fruitfulness, i.e., their usefulness for the derivation of scientific laws or theorems (Carnap 1950, *cf.* Brun 2016).[1] This line of research continues in particular in applied philosophy of science (see Schupbach 2017, pp.280-81, for examples).

Conceptual engineers have pursued both the more ambitious aim of creating new (e.g., scientific) concepts to replace extant (e.g., folk) concepts and the more modest aim of revising and improving extant stock. The past five years, however, have seen a great surge in interest in conceptual engineering with non-formal methods and modest, revisionist aims, for concepts of philosophical interest, which are expressed by words from natural language, are employed in ordinary discourse, and have traditionally been the target of conceptual analysis (see Cappelen and Plunkett 2020, for a review).[2] Such *'natural language re-engineering'* seeks to endow a lexical item that already has an established use in natural language with a new meaning or sense. Philosophers have often made such efforts without speaking of 'conceptual engineering', and it has been suggested that much traditional philosophical work is best understood as unrecognised conceptual engineering (e.g., Andow in press; Cappelen 2018).

Natural language re-engineering seems feasible as an ameliorative enterprise which both identifies *and repairs* defects, only if conceptual engineers can change the meaning of established words – if they can exercise 'meaning control' (*cf.* Ludlow 2014, p.83). This ability has been subject to a theoretical challenge. According to meta-semantic theories known as 'semantic externalism', the meaning of many words – including natural kind terms, but perhaps 'the great majority of all nouns, and … other parts of speech as well' (Putnam 1975, p.242) – is determined by external factors like the causal history of their use (Kripke 1980) or the environment in which language users live (Putnam 1975). The *'externalist challenge'* to conceptual engineering then arises from the reasonable assumption that language users – including conceptual engineers – cannot change these external factors in a way that would endow their words with a new meaning (Burgess and Plunkett 2013; Cappelen 2018; Koch 2018). This challenge has forcefully brought out the relevance of questions of control for conceptual engineering and has engendered some scepticism about its ability to go beyond identifying defects in our representational devices, and provide improved devices for actual use.[3]

[1] An earlier paradigm of conceptual engineering, neglected in current debates, is provided by the non-formal dialogical method Friedrich Waismann developed in the 1930s to identify and remove indeterminacy of meaning arising from word use specifically in philosophical contexts (Waismann 1997; *cf.* Fischer 2019).

[2] This surge in interest has been largely driven by dissatisfaction with use of the 'method of cases' to develop and test definitions, for conceptual analysis or metaphysical inquiry (*cf.* Nado 2019; Machery 2017, ch.7), as well as by a desire to make the study of concepts more societally relevant (in the wake of Haslanger 2000). Findings from 'negative' experimental philosophy suggest that the method of cases is unreliable (reviews: Mallon 2016; Stich and Tobia 2016). Psychological research suggests that most concepts lexicalised in natural language do not have the set of individually necessary and jointly sufficient application conditions that could be captured by a definition (Ramsay 1992). The move from descriptive conceptual analysis to prescriptive conceptual engineering promises to avoid these problems (Andow in press; Schupbach 2017).

[3] Cappelen (2018, pp.72-75) seems content to reduce conceptual engineering to a purely academic pursuit along the lines of some 'ideal' theorising in political philosophy. Against him, Koch (2018) argues that externalist

The present paper is the first contribution to debates about conceptual engineering to explore such questions from an empirical perspective. We will focus on the arguably most common form of natural language re-engineering actually practiced in analytic philosophy. It proceeds by giving verbal explanations which are intended to endow a familiar word with a new sense that is distinct from but related to (some of) the well-established senses in which the word is used in ordinary discourse. Typically, the new sense is introduced for specific philosophical purposes, and its explanation is invoked in developing and assessing arguments, in philosophical debates. Examples include Sally Haslanger's proposed new definition of 'woman' (Haslanger 2000, pp.42-43), David Chalmers's introduction of a new sense of 'zombie' (Chalmers 1996, pp.94-95), now recognised by the online *Oxford Dictionary*,[4] and the explanation of phenomenal senses of appearance- and perception-verbs in the philosophy of perception (e.g., Ayer, 1956, p.90, Fish, 2010, p.6, Jackson, 1977, pp.33–49; *cf.* Chisholm, 1957, pp.44–48). In its standard form, natural language re-engineering thus creates or enhances polysemy (where words have two or more distinct, but related senses) through normative semantic explanations.

> Semantic explanations of a word are *normative* when used in the relevant linguistic practice (e.g., philosophical discourse) to assess the correctness of categorization judgments, the accuracy of descriptions using the word, and the validity of inferences from utterances using the word.

The present paper will examine this standard form of conceptual engineering. We will examine whether it is feasible as an ameliorative enterprise. It is so feasible only if those who wish to use a 're-engineered' familiar word in its new sense are able to use the word in line with the new explanation, in their judgment and reasoning, across the new sense's intended range of application. For example, users have to be able to make, and take into account, inferences licensed by the new explanation and to avoid or reject inferences that are cancelled by it, wherever they or others use the word in the new sense. Otherwise, conceptual engineering will create cognitive defects (e.g., lead to systematic fallacies of equivocation that undermine our ability to reason properly), offsetting whatever advantages it might have. The cognitive competencies typically taken to involve concepts, such as language production and comprehension, verbal reasoning, perceptual categorisation, and inductive learning, are largely driven by automatic cognitive processes over which we have no direct control (see Section 2). This raises the question of 'conceptual control' (as I shall define it for present purposes):

> *Conceptual control* is the ability of thinkers to bring the exercise of their concept-involving cognitive competencies in line with relevant definitions or other normative semantic explanations.

To what extent do competent thinkers possess such control? To what extent are we able to bring our categorisation judgments, utterance interpretation, verbal reasoning, etc., in line with normative explanations provided by natural language re-engineering?

This paper will make a start on answering this question. Since natural language reengineering for philosophical purposes seeks to optimise concepts mainly for use in verbal reasoning, and typically proceeds by explaining new senses for polysemous words, we will focus on control over utterance interpretation and verbal reasoning with polysemous words, in

theories allow for the possibility of meaning change through language users' collective long-term efforts, and that such 'collective long-range control' suffices for the purposes of ameliorative conceptual engineering.

[4] Sense 1.3 in: https://www.lexico.com/definition/zombie . Last accessed 15/12/2019.

contexts where they are used in new senses. We will consider findings from psycholinguistic research on polysemy processing and findings from experimental philosophy. These findings identify '*control gaps'*, that is, parts of natural language for which thinkers are unable to bring their comprehension and verbal reasoning processes in line with the relevant normative explanations of words' new senses. This constitutes an empirical challenge to natural language re-engineering. We will develop this challenge and explore how conceptual engineers can work around it.

Section 2 will review psycholinguistic findings about polysemy processing to render the question of conceptual control – or, at any rate, our chosen part of it – empirically tractable. On this basis, Section 3 will present recent work from 'evidential' experimental philosophy that speaks to this question and identifies a philosophically relevant gap in conceptual control. Section 4 will bring out methodological consequences for natural language re-engineering. We will see that while ameliorative conceptual engineering cannot be practiced from the proverbial armchair, it is an exciting interdisciplinary project that can provide us with representational devices that help us reason better – if it engages in some detail with psycholinguistics and related branches of cognitive psychology, and engages in experimental philosophy.

2. Polysemy processing

To be able to relate questions about conceptual control to psycholinguistic research we need a psychological notion of concepts – which so far has been underutilized in debates about conceptual engineering (see Isaac 2020, for a review). A common understanding of '*concept'* in cognitive science, spelled out by Edouard Machery (2009; 2017, pp.210-212), is that of a body of information stored in long-term memory and retrieved by default, in the exercise of several higher cognitive competencies including language comprehension, perceptual categorisation, and inductive learning. That information is retrieved 'by default' here means that it is rapidly retrieved (e.g., in response to a verbal stimulus) in every context (such as, e.g., textual context), by an automatic process. A process is 'automatic' (a gradable notion) to the extent to which it is effortless (requires little attentional resources), unintentional (insensitive to the subject's goals) and unconscious (Dijksterhuis 2010). On this explication, the difference between conceptual knowledge and non-conceptual 'background knowledge' or information is made in processing terms: What distinguishes conceptual from background knowledge is that it is retrieved by default. Different kinds of information can be so retrieved, in exercising the same cognitive competence (Machery 2009). This information includes statistical information about typical and diagnostic properties of category members, captured by 'stereotypes' (or computed from 'exemplars'), as well as causal, functional, generic, and nomological information, captured by 'theories'.

This paper will focus mainly on stereotypes – also known as 'prototypes' when associated with object- or mass-nouns (Rosch 1975) and 'situation schemas' when associated with event nouns or verbs (Rumelhart 1978). We will explain what they are, identify which of them are '*conceptual structures'* representing conceptual information, and review their role in polysemy processing. This will allow us to address the two key questions concerning conceptual control about them (in this and the next section, respectively). Namely:

1) To what extent do we need to change the content or deployment of these conceptual structures, in order to bring utterance comprehension and verbal reasoning in line with normative explanations of new senses of polysemous words?[5] And:
2) To what extent are we able to effect such change?

2.1 Stereotypes as conceptual structures

As traditionally conceived, *stereotypes* represent sets of weighted features of things or events which come to mind first, and are easiest to process, when we hear words, and are diagnostic or predictive of the relevant categories (Hampton 2006). In simple cases, they can be elicited through listing and sentence-completion tasks: 'Tomatoes are___' (e.g., McRae et al. 1997). Stereotypes are built up by observation of the co-occurrence of typical properties of things and of typical components of events, in the observed physical or discourse environment (the higher the proportion of tomatoes you encounter that are red, the more strongly *red* becomes associated with *tomato*) (McRae and Jones 2013). Event nouns (Hare et al. 2009) and verbs (Ferretti et al. 2001) are associated with more complex situation schemas which include typical features of events or actions (instruments used, etc.), agents, and 'patients' acted on. For example, the situation schema associated with the verb 'S sees X' ('*seeing-schema'*) comprises typical agent features including *S looks at X*, *S knows what X is*, and *S knows that X is there* and typical patient features including *X is in front of S*, and *X is near S*.

This implicit knowledge about co-occurrence frequencies of features in the world is stored in what psychologists call 'semantic memory', which is our store of general world knowledge, as opposed to 'episodic memory' of self-experienced events (McRae and Jones 2013).[6] This world knowledge is immediately deployed in language comprehension (Elman 2009; Levinson 2000). In psycholinguistics, 'retrieval' of information is operationalised as 'activation' in priming studies.[7] Such studies have shown that single words activate stereotypical features rapidly (within 250ms) (review: Engelhardt and Ferreira 2016). Activated stereotypical features support automatic inferences from words ('tomato', 'secretary', 'S saw X').[8] These defeasible *stereotypical inferences* (e.g., the tomato referred to will have been red, the secretary female, X in front of S, etc.) are made largely irrespective of context but can typically be swiftly disregarded when they are explicitly cancelled by contextual information

[5] This is intended as an empirical question in psycholinguistic terms. While disagreeing about how different senses of polysemous words are represented (review: Eddington & Tokowicz 2015), psycholinguists think of a 'sense' as a body of information (set of features; see Sec. 2.1 below) that enters into the interpretation of relevant utterances (see, e.g., Brocher et al. 2018). This information must be activated by the word in relevant contexts but need not be retrieved by default. The normative explanations of the conceptual engineers cited above specify information (features) that need to be included in such a body of information, to achieve an intended interpretation (e.g., that the 'zombies' envisaged by Chalmers have bodies just like ours). How these conceptual engineers – or any other philosophers – theorise about 'senses' is immaterial for present purposes.

[6] Debates about the relationship between semantic and episodic memory are beyond the scope of this paper.

[7] In priming studies, participants are presented with a 'prime' word or short text and then a 'probe' word or letter string, and have to, e.g., read out the word or decide whether the string forms a word. That the prime *activates* the probe concept, i.e., makes it more accessible and likely to be used by cognitive processes (from word recognition to forward-inferencing), is inferred from shorter response times (Lucas 2000).

[8] To study automatic comprehension inferences, psycholinguists use the '*cancellation paradigm'*: Participants read or hear sentences where the expression of interest is followed by a sequel that is inconsistent with (or 'cancels') inferences the participant is hypothesised to automatically make when encountering that expression. If the hypothesised inference is made, the clash of the conclusion with the sequel will engender comprehension difficulties requiring cognitive effort. This effort is picked up by a variety of process measures including pupil dilations (Sirois & Brisson 2014), longer 'late' reading times (Clifton et al. 2007), and signature electrophysiological responses ('N400s') (Kutas & Federmeier 2011).

or clash with background knowledge (Fischer and Engelhardt 2017b). In a collaborative communicative practice (Grice 1989), such inferences are anticipated by speakers and made by hearers, in line with the 'I-heuristic' (Levinson 2000): Speakers skip mention of stereotypical features but make deviations from stereotypes explicit. In the absence of such indications to the contrary, hearers infer that the situation talked about conforms to the relevant stereotypes. Words thus trigger a host of probabilistic inferences (e.g., from 'S sees X' to *S looks at X*, *S knows X is there*, *X is near S*, etc.), in the routine comprehension process of 'stereotypical enrichment'.

As formulated by Machery (2009; 2017), the criterion for distinguishing conceptual from non-conceptual information is default activation in every context. Verbal stimuli activate all the semantic and stereotypical features associated with them, when presented in isolation, e.g., in single-word priming. Nouns activate all these features regardless of context, also when presented in a sentence context (for a review, see Giora 2003). However, the sentence context influences which parts of a more complex situation schema a verb will activate: The activation of the schema's component features depends upon fit with the thematic role (agent, patient, etc.). While the verb ('arrest') alone activates both typical agents (*cop*) and patients (*criminal*), in single-word priming experiments, sentence fragments that leave blank the agent- and the patient-role, respectively, activate only typical features of agents and patients, respectively ('She was arrested by the ___' activates *cop*, not *crook*, while 'She arrested the ___' activates *crook*, not *cop*) (Ferretti et al. 2001; *cf.* Kim et al. 2016). This suggests that we modify the present explanation of 'retrieval by default': At any rate in the context of language comprehension, we should render this as 'activation' not 'in every context' (as per Machery 2017, p.211), but 'outside all context', as in single-word priming. The moment the word is used in a discourse context, contextual cues can influence which parts of this conceptual information are activated for use in interpreting a specific utterance.

The same processes that build up situation schemas activated by individual verbs also build up situation schemas that encode more general or more specific knowledge about recurrent situations (restaurant visits, car inspections, etc.) and are activated only by combinations of words (Elman and McRae 2019). For instance, neither the noun 'mechanic', nor the verb 'check' activates things mechanics normally check, in single-word priming; but 'the mechanic checked' triggers automatic inferences to such patients (e.g., tyres and brakes, but not 'the spelling of his report'), in a self-paced reading task (Bicknell et al. 2010; *cf.* Matsuki et al. 2011). This suggests that a schema capturing information about car inspections is activated by noun and verb together, but not by either word alone. Finally, an EEG study showed that discourse context can activate a schema that is then deployed in the interpretation of a sentence that does not activate this event knowledge on its own (Metusalem et al. 2012). These schemas, which only get activated by combinations of words in more comprehensive linguistic contexts, do not qualify as capturing 'conceptual information', also on our amended explication. They represent non-conceptual 'background knowledge'.

To sum up, some stereotypes qualify as conceptual structures, while others do not.[9] Those stereotypes that are activated by individual words presented out of context (e.g., in single-word priming experiments) are conceptual structures. These include both prototypes and many, but not all, situation schemas.

[9] This conclusion is consistent with Machery's (2009) suggestion that, for the explanatory purposes of cognitive science, the notion of 'concept' should be replaced by more precise terms like 'stereotype'.

2.2 Stereotypes in polysemy processing

These conceptual structures are deployed in at least three different ways to interpret utterances with ambiguous words. Words with different unrelated meanings (*homonyms* like 'bank') activate separately represented and mutually exclusive stereotypes that compete for sustained activation (Berretta et al. 2005; Pylkkänen et al. 2006). They initially activate all these stereotypes, irrespective of context. However, the stereotype associated with the most frequently encountered sense is activated most swiftly and strongly by the verbal stimulus, irrespective of context (Giora, 2003). Contextual information then determines which of the initially activated stereotypes retains activation through integration processes including reinforcement and decay (Oden and Spira 1983), and more effortful suppression in the light of inconsistencies with contextual information (Faust and Gernsbacher 1996). The stereotype that wins this competition for sustained activation is deployed to interpret the utterance.

By contrast, words with distinct, but related senses (*polysemes*) typically activate a unified 'core representation' (Klepousniotou et al. 2012; MacGregor et al. 2015), rather than anything describable as 'representations of different senses'. Often, this 'core representation' is the stereotype associated with the most frequently encountered sense (or a sense privileged, e.g., through embodiment), which is then deployed to interpret utterances that use the word in a less frequent sense. This can happen in at least two different ways, depending upon whether that stereotype under- or over-specifies the relevant information and, accordingly, is either (a) completely or (b) only partially relevant in the utterance context. As an example, consider the interpretation of utterances that use the polysemous verb 'see' in the less frequent senses *visit*/*meet* and *know/understand*, respectively.[10]

a) *Integration Strategy (integration with background knowledge)*: Together with other words, the word of interest activates non-conceptual background knowledge that is consistent with the stereotype activated by the word alone (Bicknell et al. 2010; Matsuki et al. 2011). Where the word of interest is polysemous, this further information is used to disambiguate it. Where all conceptual information activated by the word alone is contextually relevant, disambiguation is a matter of adding further information to it. For example, together with the patient-noun 'doctor', 'see' activates the schema (or script) organising our knowledge about doctor visits. This information is compatible with that from the *seeing*-stereotype (I visually see the doctor during my visit). Therefore, both schemas remain activated and inferences from both are integrated into an interpretation of the utterance 'I saw the doctor yesterday'.[11] Similarly, such background knowledge is activated by the word together with syntactic cues including prepositions and verb aspect (Ferretti, Kutas & McRae 2007). Thus, 'went to see' activates the contextually most appropriate meeting schema (doctor visit, client meeting, social call, etc.). More generally, together with other words or syntactic constructions, 'see' activates comprehensive non-conceptual knowledge structures that organise knowledge about visits or meetings. In interpreting the overall utterance, this background knowledge is integrated with information from the seeing-schema, to obtain an interpretation that can be verbalised alternatively by using the words 'visit' or 'meet'.

b) *Retention/Suppression Strategy* (Giora 2003; Giora et al. 2014): The stereotype associated with the most frequently used sense is initially activated by the word, is retained for

[10] Senses 2 and 4 in the *Macmillan Dictionary* (*MEDAL*) (last accessed 15/12/2019): https://www.macmillandictionary.com/dictionary/british/see_1

[11] This interpretation can be rapidly amended in the light of contextual information ('He was mowing his front lawn.')

> interpreting the utterance using the word in a less frequent sense, and is ‘cut down to fit the occasion’ by suppressing its contextually irrelevant components. To interpret, for example, the metaphorical epistemic use of ‘see’ in ‘Jack sees the risk’, hearers retain the situation schema associated with the dominant visual sense of the word. This schema comprises agent-features including *S looks at X*, *S knows X is there*, and *S knows what X is*, and patient features including *X is in front of S* and *X is near S.* Hearers then suppress all contextually irrelevant component features of the schema (*S looks at X*, *X is in front of S*, etc.). Thus, only the contextually relevant epistemic agent features (*S knows X is there* and *S knows what X is*) are retained and hearers infer the intended interpretation ‘Jack knows there is a risk and knows what it is’.

These two processes interact, as integration with background knowledge and suppression of contextually irrelevant stereotype components go on in parallel, in incremental utterance interpretation.

The psycholinguistic research reviewed thus gives rise to a conception of polysemy processing which dispenses with the notion of retrieval of separate ‘entries’ for alternative senses, in a ‘mental lexicon’ (see Elman 2011, for the larger theoretical picture). It provides us with perhaps surprising answers to our first question (1 above): This research suggests that, in order to bring our comprehension and verbal reasoning processes in line with normative explanations of new senses, we do not need to change the *content* of conceptual structures associated with the polysemous word. That is, we do not need to associate a new stereotype representing conceptual information with the word, and do not need to change the component features of the stereotype associated with it, or their weights. Rather, we need to change the *deployment* of the stereotype already associated with the word. To facilitate the Integration Strategy, we need to ensure that, together with contextual cues, the word will also activate relevant non-conceptual knowledge, that this background knowledge gets integrated with information from the associated stereotype, and that both happens in all contexts to which we intend to apply the word in the new sense we explained. To facilitate the Retention/Suppression Strategy, we need to ensure that, in response to contextual cues, irrelevant components of the associated stereotype get suppressed, in all those contexts.

We now turn to our second question (2 above) and will examine to what extent competent language users are able to change how their stereotypes are deployed in language comprehension and verbal reasoning, in these ways. This will reveal an important asymmetry between the two interpretation strategies.

3. A control gap

In incremental utterance comprehension, relevant background knowledge is activated at the earliest possible moment, and swiftly integrated into the situation or discourse model on which verbal reasoning about the situation is based (Sec. 2). This suggests that we will typically be able to bring utterance interpretation and verbal reasoning in line with normative explanations of a new sense, where those cognitive processes employ the Integration Strategy. By contrast, recent findings from experimental philosophy suggest that, under certain conditions, competent thinkers lack such conceptual control, where utterance interpretation involves the Retention/Suppression Strategy. This gap in control arises from a ‘linguistic salience bias’. To establish the existence of a first philosophically relevant gap in conceptual control, we now review experimental evidence for the linguistic salience bias (Sec. 3.1) as well as textual evidence that the resulting control gap has affected philosophical argument (Sec. 3.2).

3.1 Salience bias

The ‘evidential’ research program that has emerged from ‘negative experimental philosophy’ seeks to assess the evidentiary value of philosophically relevant intuitions (for reviews, see Machery 2017; Mallon 2016; Stich and Tobia 2016). The most ambitious contributions to this program seek to do so by developing psychological explanations that trace the targeted intuitions back to specific automatic cognitive processes and help us develop ‘epistemological profiles’ that tell us under what conditions a particular process is (not) reliable (Weinberg 2015). One particularly relevant process is stereotypical enrichment (Levinson 2000), whereby automatic comprehension inferences in line with the I-heuristic (see above, Sec. 2.1) lead from verbal case descriptions to intuitions about what else is also true of the cases described. One body of work has made a start on developing an epistemological profile for this process, with a view to assessing not only intuitions (Fischer and Engelhardt 2016), but also inferences in verbal reasoning (Fischer and Engelhardt 2017a; 2017b; 2019a; 2019b; Fischer et al. 2019). While this research has been directed at other research questions, its findings speak directly to the issue of conceptual control.

Language users obviously cannot directly influence what automatic comprehension inferences are triggered by words. However, when the conclusions of such inferences clash with contextual information or background knowledge, they can be suppressed within one second and before they influence further cognition (Fischer and Engelhardt 2017b, Exp.2; *cf.* Faust and Gernsbacher 1996). The experimental work at issue has identified one set of conditions under which such suppression remains incomplete, and inappropriate inferences influence further judgment and reasoning.

A verbal stimulus activates the stereotypical features associated with the word, in its more frequent meanings or senses, regardless of context. Strength of activation depends on the non-contextual ‘salience’ of the relevant meaning or sense (Brocher et al. 2018; *cf.* Giora 2003). Such *linguistic salience* is a function of exposure frequency (how often someone hears or reads the word in this sense, rather than another), modulated by prototypicality (how good examples of the relevant category the word is deemed to stand for, when used in that sense). The Retention/Suppression Strategy (see above, Section 2.2) retains the stereotype associated with the dominant (most salient) sense of a polysemous word, to interpret utterances that employ the word in a less salient sense. This strategy requires suppression of those component features of the stereotype that are irrelevant in the contexts in which the word is applied in the less salient sense. But suppression of irrelevant components may remain partial: The stereotypical features associated with the dominant sense will be particularly strongly activated, where this sense is far more salient than all others (above). Component features that are frequently co-instantiated within the stereotype pass on activation among each other (Hare et al. 2009; McRae et. al. 2005). Where some, but not all of these ‘core components’ are contextually relevant, lateral cross-activation of irrelevant components may work in tandem with salience-based strong initial activation to render their complete suppression impossible. Schema components that are only partially suppressed continue to support stereotypical inferences that influence further cognition.

This reasoning motivates the *Salience Bias Hypothesis* (SBH) (Fischer and Engelhardt, 2019a; 2019b) that identifies a philosophically pertinent gap in our conceptual control: When

(i) one sense of a polysemous word is much more salient than all others, and
(ii) the stereotype associated with that sense is retained to interpret utterances which employ that word in a less salient sense (as per the Retention/Suppression Strategy), and
(iii) some, but not all, of the core components of the stereotype are contextually relevant,

then the less salient use of the word will trigger stereotypical inferences that are licensed only by the dominant sense and these automatic inferences will influence further judgment and reasoning, even when thinkers explicitly know they are inappropriate. That is: When conditions (i)-(iii) are met, even competent thinkers cannot help going along with automatic inferences they explicitly reject – for instance, because they are inconsistent with a normative explanation of the relevant (less salient) sense.

Four studies documented inappropriate stereotypical inferences from the verb 'to see' predicted by this hypothesis (Fischer and Engelhardt 2017a; 2017b, Exp.1; 2019a; 2019b). The visual sense is clearly the most salient sense of this verb (Fischer and Engelhardt 2019a, Table 1). The SBH thus predicts, for example, that spatial inferences (from 'S sees X' to *X is in front of S*) that are licensed by the dominant visual sense of 'see' will be prompted also by less salient epistemic uses ('Jack saw Jane's point') – and will influence further cognition, even though thinkers know they are inappropriate. In a pre-study (Fischer and Engelhardt 2019a, pp.6-7), participants indicated they were highly confident that such inferences typically *fail* to lead from true premises to true conclusions. To determine whether such inferences are made, all the same, the main studies implemented the cancellation paradigm (Fn.8) with plausibility rankings (Fischer and Engelhardt 2017a), pupillometry (Fischer and Engelhardt 2017b; 2019a), and reading time measurements with eye tracking (Fischer and Engelhardt 2019b).

In the eye tracking (pupillometry and reading time) studies, participants read sentences with concrete and abstract objects, which invited visual and epistemic interpretations of the verb, respectively. E.g.:

(1) Joe sees the problems that lie ahead.
(2) Jack sees the problems he left behind.

Evidence of spatial inferences from such epistemic uses of the verb (e.g., to *the problems are in front of Joe*) were provided by pupil dilations and longer reading times for 'spatially inconsistent' sequels ('he left behind') than for 'spatially consistent' counterparts ('that lie ahead'). Subsequent plausibility ratings provided evidence that these inferences influenced further judgment and reasoning: The sequels use familiar spatial time metaphors (behind = in the past; ahead = in the future). The future is harder to know than one's past. Accordingly, in a pre-study, participants rated the plausibility of paraphrases of purely metaphorical interpretations, and deemed paraphrases of spatially consistent sentences like (1) ('Joe knows what problems he will have in the future') less plausible than paraphrases of spatially inconsistent sentences like (2) ('Jack knows what problems he had in the past'). In the main studies, however, plausibility judgments were reversed: Spatially consistent sentences like (1) were deemed more plausible than spatially inconsistent sentences like (2). This suggests that spatial inferences continued to prevent purely metaphorical interpretation and influenced plausibility judgments. Further studies provide evidence of inappropriate inferences predicted by the Salience Bias Hypothesis, namely, from phenomenal uses of appearance verbs (Fischer and Engelhardt 2016; Fischer et al. 2019) and from a philosophical use of the noun 'zombie' (Fischer and Sytsma *ms*).

The inappropriate inferences at issue cannot be prevented by explicit marking. Where less salient meanings are associated with distinct stereotypes, riders like 'in a special sense' reinforce the activation of relevant stereotypes and help prevent them from being sidelined by dominant stereotypes, which initially receive stronger activation from the verbal stimulus (Givoni et al. 2013). Such reinforcement can prevent inappropriate inferences. To prevent the inferences posited by the Salience Bias Hypothesis, however, marking would need to reinforce suppression of components of the dominant stereotype, rather than activation of competitors. These inferences therefore cannot be prevented by highlighting that the special sense is used.

3.2 Loss of conceptual control in philosophical argument

The gap in conceptual control thus identified has arguably affected philosophical argument deploying 're-engineered' terms from ordinary discourse. For instance, philosophers of perception have engaged in natural language re-engineering by introducing a new phenomenal sense of appearance- and perception-verbs. In this sense, which is not recognized by the major dictionaries,[12] these verbs are used purely to describe the perceiver's subjective experience and are stipulated to lack all factive and doxastic implications (e.g., Ayer, 1956, pp.98-104, Fish, 2010, p.6, Jackson, 1977, pp.33–49; *cf.* Chisholm, 1957, pp.44–48). To say that S 'sees' an F, in this sense, simply is to say that S has a subjective experience like that of seeing an F; it does not imply that there actually is an F (or no F) around to be seen, or that S makes any judgment, not even about the character of their current experience.

The three conditions for salience bias are met: As per condition (i), the visual sense, e.g., of 'see' is clearly dominant, while its phenomenal sense is even less salient than its epistemic sense (Fischer and Engelhardt 2019a, Table 1). Fischer and Engelhardt (op. cit.) suggest that, as per condition (ii), the metaphorical phenomenal sense is interpreted with the Retention/Suppression Strategy: The situation schema associated with the visual sense of 'see' is retained to build a situation-model that instantiates the schema with specific patient-role fillers (e.g., a dagger). This model contains a set of phenomenal features as a component, and these features are attributed to the target experience, in a variant of the common 'feature transfer' approach of metaphor interpretation (Ortony 1993). However, what it is like to see something is strongly associated with spatial features of the schema associated with the dominant visual use of 'S sees X', as evidenced by embodied cognition effects associated with visual metaphors (Lakoff 2012). These frequently co-instantiated core schema components thus exchange activation. Continued activation of phenomenal features thus maintains activation of factive/spatial features (*There actually is an X there, near S*). When, as per condition (iii), some of these core components are contextually irrelevant (as in talk about hallucinations), their activation cannot be suppressed, and the Salience Bias Hypothesis predicts that even competent thinkers will go along with inappropriate factive/spatial inferences from phenomenal uses – even when this special use is explicitly marked (Section 3.1).

A case in point is the 'argument from hallucination' frequently adduced against naïve realism or for sense-datum theories (e.g., Ayer 1956; Smith 2002). This classic statement explicitly marks its use of the special phenomenal sense, which is subsequently explained in some detail (Ayer 1956, pp.98-104):[13]

> 'Let us take as an example Macbeth's visionary dagger [...] There is an obvious [perceptual] sense in which Macbeth did not see the dagger; he did not see the dagger for the sufficient reason that *there was no dagger there* for him to see. There is another [viz., phenomenal] sense, however, in which it may quite properly be said that *he did see a dagger;* to say that he saw a dagger is quite a natural way of describing his experience. *But still not a real dagger; not a physical object... If we*

[12] See, e.g., *MEDAL* (Fn.9), *Oxford Dictionaries* (https://www.lexico.com/definition/see), or *Princeton WordNet*. The *Oxford English Dictionary*, 'see' sense 11a, comes close but covers only non-perceptual experience (last accessed 15/12/2019).

[13] In explaining his intended phenomenal use of 'see', Ayer subsequently proposes the neologisms 'have in sight' (Ayer 1956, pp. 100 and 104) and 'seem to see' (pp.101-104), to explicitly mark this use.

> *are to say that he saw anything, it must have been* something that was accessible to him alone… *a sense-datum*.’ (Ayer 1956, 90; my italics).

The second half of the argument then generalises to all cases of visual perception.

The following reconstruction remains as close to the text as possible and builds the intended deductive argument from the italicised text:

(1) ‘There was no [real] dagger there.’
(2) ‘Macbeth did see a dagger.’

To deductively infer that Macbeth did not see a real dagger (‘But still not a real dagger’), we need an implicit assumption:

[3] If Macbeth saw a real dagger, there was a real dagger there. By (1) & [3] with *modus tollens*:
(4) ‘Macbeth did not see a real dagger.’
[5] Macbeth did not see any other physical object. By (4) & [5]:
(6) ‘Macbeth did not see a physical object.’ By (6):
(7) ‘If… he saw anything,’ he saw a non-physical object, a ‘sense-datum.’ By (7) & (2):
(8) Macbeth saw a sense-datum.

Premise (2) explicitly uses the verb in the phenomenal sense that has been stipulated to lack factive/spatial implications. *Pace* [3], that Macbeth ‘saw’ a dagger in *this* sense does *not* imply there was a dagger there. Indeed, the phenomenal sense is to describe what the subjective experience is like, and Macbeth’s subjective experience is just like that of seeing a real, physical dagger, by philosophical assumption. Therefore, *pace* (4) and (6), Macbeth *can* be said to ‘see’, in *this* sense, a real, physical dagger and hence a physical object – but not, e.g., an unreal, translucent dagger (his subjective experience is not like that). These assumptions and conclusions are only true if their use of ‘see’ is interpreted in line with the ordinary, perceptual sense, rather than the phenomenal sense explicitly adopted for (2). But if the verb is used in different senses, the final inference commits a fallacy of equivocation. Either way, Ayer’s reasoning fails to consistently abide by the relevant explanation of the phenomenal sense of perception-verbs.

Since the different senses of ‘see’ are explicitly marked in the passage, this failure cannot be a mere slip. Rather, this failure illustrates the lack of conceptual control predicted by the Salience Bias Hypothesis. Arguably, Ayer’s stated argument is an ex-post rationalization of the intuitive line of thought (*cf.* Smith 2002, p.194): ‘Macbeth saw a dagger. So there was a dagger. But there was no physical dagger around. So he must have seen a non-physical “dagger” (a dagger-like sense datum).’ The first step is the predicted factive/spatial inference from a phenomenal use. The resulting conditional is presupposed, in the shape of [3] above, by the explicit argument. The salience bias thus leads proponents of the argument to make an inappropriate factive/spatial inference from the phenomenal use of ‘see’ and to presuppose the conclusion in their further reasoning. Due to the bias, they cannot help going along with automatic inferences that are cancelled by their own normative explanations of the phenomenal sense.

Further philosophical arguments that arguably illustrate such persistent lack of conceptual control include the argument from illusion (Ayer 1956; Robinson 1994; Smith 2002) and David Chalmers’ (1996, pp.94-96) ‘zombie argument’ (see Fischer et al. 2019, and Fischer and Sytsma *ms*, respectively). Published statements of the Salience Bias Hypothesis (Fischer and Engelhardt 2019a; 2019b; Fischer et al. 2019) state it applies to high-frequency words, and the supporting studies examine it for verbs, which play a particularly fundamental role in incremental sentence comprehension, in verb-medial languages like English (Melinger and Mauner 1999; Tanenhaus and Carlson 1989). However, a subsequent study examined the extension of the hypothesis to low-frequency nouns and provided experimental evidence of contextually inappropriate stereotypical inferences from the low-frequency noun 'zombie' (Fischer and Sytsma *ms*). This suggests that salience bias manifests itself more widely, and that the resulting lack of conceptual control may vitiate several important philosophical arguments.

4. Consequences for conceptual engineering

Our engagement with psycholinguistics and experimental philosophy has two key upshots for conceptual engineering. First, in order to bring people's comprehension and verbal reasoning processes in line with normative explanations of the new senses they introduce, conceptual engineers do not necessarily need to change the *content* of conceptual structures associated with the polysemous word (the component features of stereotypes or their weights); it may suffice to change the way those structures are *deployed* (Section 2). This appears to render the standard form of natural language re-engineering feasible in the face of a difficulty: Since stereotypical associations are (sluggishly) responsive to observed co-occurrence frequencies in the physical and discourse environments, they can only be changed – over time – by changing these environments. The mere introduction of new senses of words through normative verbal explanations that are relevant only for specific (e.g., research) contexts is unlikely to achieve such change (*cf.* below). The discovery that changes in the deployment of conceptual structures may suffice to provide us with conceptual control over new senses then holds out a promise: The verbal explanations of new senses offered by natural language re-engineering might provide us with linguistic tools that help us think better, without us having to change the world first. Our second key finding, however, revealed limitations of this promise: Under certain conditions, the salience bias prevents appropriate deployment of stereotypical information and thus creates a philosophically relevant gap in competent thinkers' conceptual control (Section 3). The identification of this control gap creates an empirical challenge to natural language re-engineering (Sec. 4.1). However, the empirical findings simultaneously suggest ways in which conceptual engineers can address this challenge and work around the gap (Sec. 4.2).

4.1 The empirical challenge

Natural language re-engineering introduces new senses of familiar words through normative explanations. Where this happens, thinkers engaged in utterance interpretation and verbal reasoning with such a new sense have to be able to do two things: They need to be able, first, to make, and take into account, inferences that are licensed by the normative explanations of the new sense and, second, to avoid relying on inferences that are inconsistent with (cancelled by) these explanations.

The Salience Bias Hypothesis pinpoints a gap in such conceptual control – potentially, one of several such gaps. It identifies conditions under which competent thinkers lack the second ability: Under these conditions, competent thinkers are unable to suppress automatic inferences which are licensed only by the dominant sense of a polysemous word, even where these are triggered by a use of the new sense and are cancelled by the normative explanations of this new sense. This will happen where the stereotype associated with a dominant familiar sense is only partially contextually relevant and functional for interpreting uses of the word in its new sense, with the Retention/Suppression Strategy, while the familiar dominant sense is far more salient than all the other senses (Sec. 3.1).

This situation will often arise in natural language re-engineering and render it vulnerable to salience bias. In ordinary discourse, many words have dominant senses that are far more salient than all others. Natural language reengineering in philosophy tends to introduce quite special senses of familiar words, which are introduced for specific philosophical purposes (for instance, in our main example, to talk just about perceivers' subjective experience). Such new senses are likely to find uptake only in a few further contexts, while even professional philosophers using them will continue to be immersed in ordinary discourse that exposes them to the dominant sense as frequently, and in similar situations, as before, and more frequently to it than to any other sense. This has two consequences: First, the stereotype associated with the dominant sense of the word is unlikely to change much in terms of weights of constituent features. Second, in particular for high-frequency words, the introduction of the new special sense is unlikely to greatly diminish any pronounced salience imbalance between the dominant and the other senses, even for a philosophical audience. Where polysemy processing is governed by the Retention/Suppression strategy, salience bias will result and vitiate the efforts of natural language reengineering.

The bias will undermine the point of such re-engineering. The bias will lead thinkers to systematically rely on inferences from uses of the new sense that are licensed only by the dominant sense and are inconsistent with the conceptual engineer's explanations. Thinkers will do so, for instance, in the interpretation of philosophical texts and in philosophical argument. This will systematically lead to fallacies of equivocation like the fallacy identified in the influential argument from hallucination (Section 3.2). Where it is affected by salience bias, natural language re-engineering is thus set to reduce, rather than enhance, our ability to reason properly in philosophy. The challenge to this project is to avoid this pitfall – and similar pitfalls created by other gaps in conceptual control.

While this paper reviewed findings that pinpoint a gap in thinker's ability to exercise conceptual control over the deployment of stereotypes, extant research has also identified such a gap for theories, the other generally recognized kind of conceptual structure. The acquisition of new theories, e.g., in science education merely suppresses, but does not erase prior naïve theories (Shtulman and Valcarel 2012), which may continue to influence cognition. For example, Arvid Guterstam and colleagues (2019) provided experimental evidence that folk-physical judgments are influenced by an implicit 'extramissionist' theory of vision, which 95% of participants explicitly reject. According to this theory, eyes emit invisible beams of force-bearing energy that act on objects of sight, in visual perception. This implicit theory influenced judgments about the mechanical forces acting on objects: When tilted objects were pictured as being viewed by someone, the threshold angle at which participants judged the object to fall over changed, corresponding to a force of one hundredth of a Newton acting on the object against the direction of tilt, in the direction of gaze. The authors suggest the implicit theory is stored as a schematic model that is automatically deployed in monitoring the gaze of other agents – and is therefore retrieved by default whenever certain perceptual cues are present (whenever we perceive sighted agents) – and also in producing and understanding culturally influential metaphors. The implicit model is articulated by philosophical theories (see Gross 1999, for a review). Automatic implicit models built around spatial schemas – like the Cartesian Theatre model (Dennett 19991) – may influence more generally the judgments and inferences of thinkers who explicitly reject the philosophical theories that articulate them. Further research is likely to reveal further gaps in conceptual control.

4.2 Three rules for natural language re-engineering

An empirically grounded understanding of how gaps in conceptual control arise can help conceptual engineers to work around the gaps identified. The research we reviewed motivates three rules of thumb that allow conceptual engineers to work around the gap arising from the salience bias. These rules guide the choice of word that is to be given a new sense, to convey certain information (in that new sense). Applying these rules requires some empirical work and gives experimental philosophy yet another role to play in conceptual engineering, in addition to the actual and possible contributions already discussed in the literature (e.g., Fisher 2015, Machery 2017, Nado 2019; Schupbach 2017, Shepherd & Justus 2015),

Rule 1 – Where possible, introduce senses that can be interpreted with the Integration Strategy, rather than the Retention/Suppression Strategy.

The empirical research reviewed revealed an important asymmetry between two polysemy processing strategies: Where the stereotypical information activated by a polysemous word is consistent with contextual information, and disambiguation and utterance interpretation 'only' require integrating these different bodies of information, language users are good at making appropriate inferences and avoiding inappropriate inferences. By contrast, where initially activated information is partially inconsistent with contextual information, and disambiguation and utterance interpretation require suppression of contextually irrelevant information, language users are unable to avoid inappropriate inferences, under certain conditions.

This asymmetry should inform the choice of word to convey the intended information, in a new sense: To minimize the risk of creating gaps in conceptual control, conceptual engineers should prefer familiar words that, in their dominant sense, underspecify the information to be carried by them in the new sense, and introduce the new sense through an explanation that specifies, or is suitable to activate, the further information to be conveyed. For example, some subordinate senses of 'woman' (as *significant other* and *cleaning woman*) are clearly consistent with the conceptual information carried by the most salient sense (*female adult*)[14] and are interpreted by deploying background knowledge (about romantic relationships or household employment) that is contextually relevant, in line with the Integration Strategy (Sec. 2.2). Similarly, the intended new sense of 'systemically oppressed female adult' (*cf.* Haslanger 2000) can be conveniently attached to the established word 'woman', because (outside matriarchal societies) the stereotype associated with the word's most salient sense is arguably consistent with the new sense which merely adds information to that stereotypical information. This new sense is best introduced through explanations that specify relevant forms of oppression in a way that facilitates activation of the pertinent background knowledge (about oppressive social practices).

More generally, the conceptual engineer should determine the stereotype associated with a candidate word, by administering listing and typicality rating tasks (McRae et al. 1997), tentatively phrase normative explanations of the word's new sense, and check whether these explanations cancel component features of the elicited stereotype. In particular where explanations cancel features that were rated highly typical and high ratings correlate with those of other features, which remain relevant, the conceptual engineer should look for a different word to attach the intended information to, through introduction of a new sense.

Rule 2 – Where you cannot avoid reliance on the Retention/Suppression Strategy, avoid words with a pronounced salience imbalance.

Where no familiar word underspecifies the intended information at a helpful level, natural language re-engineering needs to fall back on words that over-specify the intended information.

[14] See http://wordnetweb.princeton.edu/ for frequency information.

The intended interpretation then requires suppression of some components of the stereotype associated with the dominant sense. Suppression will remain incomplete, and inappropriate inferences persist, where this sense is far more salient than all others (Sec. 3.1).

To avoid words with such a clearly dominant sense, conceptual engineers need to collect salience information. Salience is a function of exposure frequency, modulated by prototypicality. Exposure frequency cannot be directly measured. Common proxy measures are occurrence frequencies in corpora and familiarity ratings (Giora 2003). Accordingly, conceptual engineers can consult frequency information from corpora like the *Corpus of Contemporary American English* (*COCA*), or from *Princeton WordNet*. For a more rigorous examination, they can ask independent annotators to classify as uses of different senses occurrences of the word in samples drawn at random from corpora of interest (such as the 1000-sentence samples of 'see'-sentences used by Fischer and Engelhardt 2017a).[15] The other component of salience, prototypicality, can be studied through sentence-completion tasks (Chang 1986). Participants are asked to produce sentences with the word of interest, and the resulting corpus is annotated as described. So far, salience bias effects have been documented for words where the dominant sense accounted for at least two thirds of all uses in consulted and produced corpora (Fischer and Engelhardt 2019a on 'see'; Fischer and Sytsma *ms* on 'zombie'). An upper bound at which salience imbalances do not yet support inappropriate inferences remains to be identified.

Rule 3 – Where you cannot avoid using a word with a pronounced salience imbalance, ensure that the relevant stereotypical implications are not dependent upon, or typically co-instantiated with, irrelevant stereotypical implications.

This will reduce lateral cross-activation between contextually relevant and irrelevant components of the initially activated stereotype. However, even so, irrelevant components are likely to be completely suppressed only where the irrelevance of those component features is made contextually highly salient, for example, through countervailing stereotypical implications from context words (Fischer and Engelhardt 2019b, pp. 81-83 and 85-88). Where interpretation of a new sense requires us to suppress component features of the stereotype associated with a clearly dominant sense, the new sense requires careful handling and should only be used together with expressions that support their suppression by having implications that cancel the unwanted stereotypical implications from the dominant sense. The introduction of the new sense should then flag this issue.

These three rules of thumb allow the standard approach of natural language re-engineering to work around the salience bias and introduce new special senses of words that can be successfully used for specific research purposes, while those words continue to be used in their dominant senses, in most contexts. By contrast, linguistic activists with a socio-political agenda could work around salience bias by seeking to promote wider uptake of their new sense to a point where it ties with, or becomes, the dominant sense. A model would be provided by political activism that led to the suffix '-phobic' being used even more frequently to designate negative attitudes towards social groups ('homophobic', 'xenophobic', etc.) than to speak about medical health conditions ('claustrophobic', etc). First findings (Fischer and Engelhardt 2017b, Exp.2) suggest that while this change in relative frequency did not change the associated stereotype (namely, the association of '-phobic' with a mental health condition), and uses of

[15] For a review of relevant methods from corpus linguistics, with philosophical applications, see Sytsma et al. 2019.

the new sense ('S is homophobic') still trigger inappropriate stereotypical inferences (to *S has a mental health condition*), these inferences do not influence further judgment and reasoning.

5. Conclusion

The most common form of conceptual engineering, natural language re-engineering, introduces new senses of familiar words through normative explanations. To benefit from their introduction, language users must be able to bring utterance interpretations and verbal reasoning in line with these explanations. But we cannot simply presuppose that language users possess such conceptual control. This paper empirically identified a philosophically relevant gap in conceptual control – probably the first of several such gaps. If natural language re-engineering is to enhance, rather than reduce, our ability to reason properly, it needs to identify, and work around, these gaps. To be able to do so, it needs to take into account findings from cognitive psychology, and especially psycholinguistics, and to engage in experimental philosophical work. The present paper got this interdisciplinary research program started by examining how salience bias creates a gap in conceptual control and how conceptual engineers can work around this gap – and actually improve our ability to reason with words.

Acknowledgments

For helpful comments on previous drafts, I thank James Andow, Manuel Gustavo Isaac, Steffen Koch, and anonymous reviewers.

References

Andow, J. (in press). Intuitions about cases as evidence (about how we should think). *Inquiry.*

Ayer, A.J. (1956). *The Problem of Knowledge*. Repr. 1990. London: Penguin.

Beretta, A., Fiorentino, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: an MEG study. *Cognitive Brain Research*, 24, 57-65.

Bicknell, K., Elman, J.L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language, 63,* 489–505.

Brocher, A., Koenig, J.-P., Mauner, G., & Foraker, S. (2018). About sharing and commitment: the retrieval of biased and balanced irregular polysemes. *Language, Cognition and Neuroscience*, 33, 443-466.

Brun, G. (2016). Explication as a method of conceptual re-engineering. *Erkenntnis*, 81, 1211–1241.

Burgess, A. & Plunkett, D. (2013). Conceptual ethics (I). *Philosophy Compass*, 8, 1091–101.

Cappelen, H. (2018). *Fixing Language: An Essay on Conceptual Engineering*. Oxford: OUP

Cappelen, H. & Plunkett, D. (2020). A guided tour of conceptual engineering and conceptual ethics. In their (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: OUP

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: University of Chicago Press

Chalmers, D. (1996). *The Conscious Mind*. Oxford: OUP.

Chang, T. M. (1986). Semantic memory: Facts and models. *Psychological Bulletin*, 99, 199–220.

Chisholm, R. (1957). *Perceiving*. Ithaca: Cornell UP.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R.P.G. van Gompel et al. (eds.), *Eye Movements. A Window on Mind and Brain* (pp.341–371), Elsevier

Dennett, D. (1991). *Consciousness Explained*. Boston: Little, Brown & Co.

Dijksterhuis, A. (2010). Automaticity and the unconscious. In Fiske, S.T., Gilbert, D.T., & Lindzey, G. (eds.), *Handbook of Social Psychology* (p. 228–267). Wiley

Eddington, C. M., & Tokowicz, N. (2015). How meaning similarity influences ambiguous word processing: The current state of the literature. *Psychonomic Bulletin & Review*, 22, 13–37.

Elman. J.L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognition*, *33,* 547–582.

Elman. J.L. (2011). Lexical knowledge without a lexicon? *The Mental Lexicon*, 6, 1-33.

Elman, J.L. & McRae, K. (2019). A model of event knowledge. *Psychological Review*, 126, 252-291

Engelhardt, P. E. & Ferreira, F. (2016). Reaching sentence and reference meaning. In P. Knoeferle, P. Pyykkonen, & M.W. Crocker (eds.), *Visually Situated Language Comprehension* (pp. 127–150). Amsterdam: John Benjamins.

Faust, M.E., & Gernsbacher, M.A. (1996). Cerebral mechanisms for suppression of inappropriate information during sentence comprehension. *Brain and Language, 53,* 234-259.

Ferretti, T. R., Kutas, M., & McRae, K. (2007). Verb aspect and the activation of event knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33,* 182–196.

Ferretti, T., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language, 44*, 516-547.

Fischer, E. (2019). Linguistic self-explication and psycholinguistic experiments. Redeveloping Waismann's approach. In D. Makovec & S. Shapiro (eds.), *Friedrich Waismann: The Open Texture of Analytic Philosophy* (pp. 211-241). Basingstoke: Palgrave

Fischer, E. & Engelhardt, P. E. (2016). Intuitions' linguistic sources: Stereotypes, intuitions, and illusions. *Mind & Language,* 31, 67–103.

Fischer, E. & Engelhardt, P. E. (2017a). Diagnostic experimental philosophy. *Teorema*, 36(3), 117–137.

Fischer, E. & Engelhardt, P. E. (2017b). Stereotypical inferences: Philosophical relevance and psycholinguistic toolkit. *Ratio*, 30, 411–442.

Fischer, E. & Engelhardt, P. E. (2019a). Lingering Stereotypes: Salience bias in philosophical argument. *Mind and Language*. 2019; 1-25. https://doi.org/10.1111/mila.12249

Fischer, E. & Engelhardt, P. E. (2019b). Eyes as windows to minds: Psycholinguistics for experimental philosophy. In E. Fischer & M. Curtis (eds.), *Methodological Advances in Experimental* Philosophy (pp. 43–100). London: Bloomsbury.

Fischer, E., Engelhardt, P. E., Horvath, J. & Ohtani, H. (2019). Experimental ordinary language philosophy: A crosslinguistic study of defeasible default inferences. *Synthese*. https://doi.org/10.1007/s11229-019-02081-4

Fischer, E., & Sytsma, J. (ms). *Zombie intuitions*. University of East Anglia.

Fish, W. (2010). *Philosophy of Perception*. London: Routledge.

Fisher, J. C. (2015). Pragmatic experimental philosophy. *Philosophical Psychology, 28*, 412–433.

Giora, R. (2003). *On Our Mind. Salience, Context, and Figurative Language.* Oxford: OUP.

Giora, R., Raphaely, M., Fein, O., & Livnat, E. (2014). Resonating with contextually inappropriate interpretations: The case of irony. *Cognitive Linguistics*, *25,* 443-455.

Givoni, S., Giora, R., & Bergerbest, D. (2013). How speakers alert addressees to multiple meanings. *Journal of Pragmatics*, *48,* 29-40.

Gross, C.G. (1999). The fire that comes from the eye. *Neuroscientist*, 5, 58–64.

Guterstam, A., Keana, H.H., Webba, T.W., Keana, F.S., & Grazianoa, M.S.A. (2019). Implicit model of other people's visual attention as an invisible, force-carrying beam projecting from the eyes. *Proceedings of the National Academy of Sciences*, 116, 328–333.

Hampton, J. (2006). Concepts as prototypes. In B.H. Ross (ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory* (pp. 79–113). Amsterdam: Elsevier.

Hare, M., Jones, M., Thomson, C., Kelly, S., & McRae, K. (2009) Activating event knowledge. *Cognition, 111,* 151-167.

Haslanger, S. 2000. Gender and race. (What) are they? (What) do we want them to be? *Nous*, 34, 31–55.

Isaac, M.G. (2020). How to conceptually engineer conceptual engineering. *Inquiry*. https://doi.org/10.1080/0020174X.2020.1719881

Jackson, F. (1977). *Perception*. Cambridge: CUP.

Kim, A.E., Oines, L.D., Sikos, L. (2016). Prediction during sentence comprehension is more than a sum of lexical associations: the role of event knowledge. *Language, Cognition, and Neuroscience*, *31,* 597-601.

Klepousniotou, E., Pike, B., Steinhauer, K., & Gracco, V. (2012). Not all ambiguous words are created equal: an EEG investigation of homonymy and polysemy. *Brain and Language*, 123, 11-21.

Koch, S. (2018). The externalist challenge to conceptual engineering. *Synthese*, doi:10.1007/s11229-018-02007-6

Kripke, S. A. (1980). *Naming and Necessity*. Oxford: Blackwell

Kutas, M., & Federmeier, K.T. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62,* 621-647.

Lakoff, G. (2012). Explaining embodied cognition results. *Topics in Cognitive Science*, 4, 773–785.

Levinson, S.C. (2000). *Presumptive Meanings. The Theory of Generalized Conversational Implicature*, Cambridge, Mass.: MIT Press.

Lucas, M. (2000). Semantic priming without association: a meta-analytic review. *Psychonomic Bulletin and Review*, *7,* 618–630.

Ludlow, P. (2014). *Living words: Meaning underdetermination and the dynamic lexicon.* Oxford: OUP

MacGregor, L.J., Bouwsema, J. & Klepousniotou, E (2015). Sustained meaning activation for polysemous but not homonymous words: Evidence from EEG. *Neuropsychologia*, 68, 126 - 138.

Machery, E. (2009). *Doing without Concepts.* Oxford: OUP

Machery, E. (2017). *Philosophy within its Proper Bounds*. Oxford: OUP

Mallon, R. (2016). Experimental philosophy. In H. Cappelen, T. Szabo Gendler, & J. Hawthorne (eds.), *Oxford Handbook of Philosophical Methodology* (pp. 410-433). Oxford: OUP.

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., & McRae, K. (2011). Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *37,* 913–934.

McRae, K., Ferretti, T.R., & Amyote, I. (1997). Thematic roles as verb-specific concepts. *Language and Cognitive Processes, 12*, 137-176.

McRae, K., Hare, M., Elman, J. L., & Ferretti, T. R. (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, *33,* 1174-1184.

McRae, K., & Jones, M. (2013). Semantic memory. In D. Reisberg (ed.), *Oxford Handbook of Cognitive Psychology*, Oxford: OUP.

Melinger, A., & Mauner, G. (1999). When are implicit agents encoded? Evidence from crossmodal priming. *Brain and Language, 68,* 185-191.

Metusalem, R., Kutas, M., Urbach, T.P., Hare, M., McRae, K., & Elman, J.L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66,* 545-567.

Nado, J. (2019). Conceptual engineering via experimental philosophy. *Inquiry*. https://doi.org/10.1080/0020174X.2019.1667870

Oden, G.C., & Spira, J.L. (1983). Influence of context on the activation and selection of ambiguous word senses. *Quarterly Journal of Experimental Psychology, 35A,* 51–64.

Ortony, A. (1993). The role of similarity in similes and metaphors. In A. Ortony (Ed.), Metaphor and thought (2nd ed., pp. 342–356). Cambridge: CUP.

Putnam, H. (1975). The meaning of 'meaning'. In his, *Philosophical Papers*. *Mind, Language and Reality* (Vol. 2, pp. 215–271). Cambridge: CUP

Pylkkänen, L., Llinás, R., & Murphy, G. L. (2006). The representation of polysemy: MEG evidence. *Journal of Cognitive Neuroscience*, 18, 97-109.

Ramsay, W. (1992). Prototypes and conceptual analysis. *Topoi*, 11, 59–70.

Robinson, H. (1994). *Perception*. London: Routledge.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, 7, 532–547.

Rumelhart. D.E. (1978). Schemata: The building blocks of cognition. In R. Spiro, B. Bruce, & W. Brewer (eds.), *Theoretical Issues in Reading Comprehension*. Hillsdale, N.J.: Erlbaum.

Schupbach, J.N. (2017). Experimental Explication. *Philosophy and Phenomenological Research*, 94, 672-710

Shepherd, J., & Justus, J. (2015). X-Phi and Carnapian explication. *Erkenntnis, 80*, 381–402.

Shtulman, A. & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, 124, 209–215.

Sirois, S. & Brisson, J. (2014). Pupillometry. *WIREs Cognitive Science*, 5, 679–692.

Stich, S., & Tobia, K. (2016). Experimental philosophy and the philosophical tradition. In J. Sytsma & W. Buckwalter (eds.), *Blackwell Companion to Experimental Philosophy* (pp.5-21). Malden: Wiley Blackwell

Smith, A.D. (2002). *The Problem of Perception*. Cambridge, Mass: Harvard UP.

Sytsma, J., Bluhm, R., Willemsen, P., & Reuter, K. (2019). Causation attributions and corpus analysis. In E. Fischer & M. Curtis (eds.), *Methodological Advances in Experimental Philosophy* (pp. 209–238). London: Bloomsbury.

Tanenhaus, M.K., & Carlson, G.N. (1989). Lexical structure and language comprehension. In W. Marslen-Wilson (ed.), *Lexical Representation and Process* (pp. 529-561). Cambridge, MA: MIT Press.

Waismann, F. (1997). *Principles of Linguistic Philosophy*. 2nd ed. Ed. by R. Harré. London: Macmillan.

Weinberg. J. (2015). Humans as instruments, on the inevitability of experimental philosophy. In: E. Fischer & J. Collins (eds.), *Experimental Philosophy, Rationalism, and Naturalism* (pp. 171-187). London: Routledge.