*Croatian Journal of Philosophy*
Vol. XXII, No. 64, 2022
https://doi.org/10.52685/cjp.22.64.6
Received: November 20, 2021
Accepted: April 29, 2022

# *Non-Stupidity Condition and Pragmatics in Artificial Intelligence*

BOJAN BORSTNER and NIKO ŠETAR
*University of Maribor, Maribor, Slovenia*

*Symbol Grounding Problem (SGP) (Harnad 1990) is commonly considered one of the central challenges in the philosophy of artificial intelligence as its resolution is deemed necessary for bridging the gap between simple data processing and understanding of meaning and language. SGP has been addressed on numerous occasions with varying results, all resolution attempts having been severely, but for the most part justifiably, restricted by the Zero Semantic Commitment Condition (Taddeo and Floridi 2005). A further condition that demands explanatory power in terms of machine-to-human communication is the Non-Stupidity Condition (Bringsjord 2013) that demands an SG approach to be able to account for plausibility of higher-level language use and understanding, such as pragmatics. In this article, we undertake the endeavour of attempting to explain how merging certain early requirements for SG, such as embodiment, environmental interaction (Ziemke 1998), and compliance with the Z-Condition with symbol emergence (Sun 2000; Tangiuchi et al. 2016, etc.) rather than direct attempts at symbol grounding can help emulate human language acquisition (Vogt 2004; Cowley 2007). Along with the presumption that mind and language are both symbolic (Fodor 1980) and computational (Chomsky 2017), we argue that some rather abstract aspects of language can be logically formalised and finally, that this melange of approaches can yield the explanatory power necessary to satisfy the Non-Stupidity Condition without breaking any previous conditions.*

**Keywords:** Artificial intelligence; symbol grounding; pragmatics; language; computationalism.

## 1. *Introduction*

Artificial intelligence is as hot a topic as any during the last few decades, with debates on it ranging from AI ethics to its development to whether it is achievable at all. Currently, a lot of progress is being made in the development and production of neural networks and machine learning systems, yet it would seem that those systems are still not much more than just increasingly sophisticated software on increasingly sophisticated hardware. The key difference between them and artificial intelligence is, well, intelligence. Here we reach a whole different debate: what exactly does it mean to be intelligent? There is an abundance of answers, or at least attempts at answering, but let us make it simple and agree that intelligence is inextricably linked with understanding – therefore, in order to be intelligent, a machine has actually to understand the data it is processing, and not just merely process it. And that is only the beginning in the long process aimed at achieving human-like intelligence or even superintelligence.

In this article, we will overview one of philosophers' favourite approaches to making AIs understand their data – solving the Symbol Grounding Problem, which we shall introduce in the next section. We will study several proposed solutions, cherry-picking certain elements to comprise a strategy with a decent chance of success. Afterwards we will address the issue of whether any approach to grounding has the explanatory power as to how human-level artificial intelligence could be achieved and explain how this may be within our reach if we explain how complex features of language such as speech acts, metaphors, and humour may be grounded in simpler features (non-connoted sentences, words) that are in turn grounded directly.

## 2. *Symbol grounding problem*

The Symbol Grounding Problem was first formulated by Stevan Harnad (1990) and is derived from John Searle's (1980) Chinese Room thought experiment. Searle describes a room containing a vast number of monolingual resources in Mandarin Chinese, from dictionaries to encyclopaedias and novels. There is also an English-speaking man in this room who has no knowledge whatsoever of the Chinese language or writing. Next, we insert a paper page with a number of questions in Chinese that our man in the room must answer. Searle claims that with enough time (or processing power) he can find corresponding patterns of symbols in the available resources and copy the symbols that follow the question mark until the end of the sentence or paragraph. Then he outputs the paper with what are likely perfectly correct answers. However, through this process, the man in the room never understood a single Chinese symbol he was looking up or copying and had no idea what the input questions or his own output answers were. This is analogous to how computers process data: they operate based

on an algorithmic script. When they receive an input X, they 'look for' a part of their code that says something like "if X then Y," and output Y accordingly. Thus, when a command is typed into a computer and the computer performs this command, it does so without understanding what it just did, what the input meant, or what the output meant.

Harnad (1990) says that symbol grounding problem comes in two forms, the first of which is not unlike learning Chinese as a second language, using only a monolingual Chinese dictionary, which is a rather difficult task. The second form is like trying to learn Chinese as one's first language using only such a dictionary – an impossible task. Since symbol grounding that we are talking about when speaking of AI is essentially a form intended to ground a first language, such learning is impossible. What we need are external (real-world) referents to which we can relate the symbols we are manipulating.

### 2.1. *Approaches and conditions*

Harnad (1990) himself proposes a representationalist approach towards symbol grounding. The approach is based on the notion that the distal objects are projected onto the perceiver's sensory surfaces when they are perceived via any available means, be it sight, hearing, touch, or any other sensory tool, and is drawn from the work of Shepherd and Cooper (1982). Harnad dubs these projections as representations and defines several kinds of representations that manifest throughout the process of transcription of distal objects into symbols within one's mind as a symbolic system (for further details on mind as a symbolic system see Fodor 1980). When we are exposed to a particular referent in the outside world, an iconic representation of it is created; a group of referents with similar properties, in turn, yields categorical representation. Two cognitive mechanisms manipulate those representations: discrimination allows us to distinguish between different categories, as well as different tokens within a category; identification lets us recognise something in the outside world as a token belonging to a category. When related to a particular symbol (spoken/written word or such), the symbolic representation of a token or a category is formed. Regier (1992) attempts to recreate a similar bottom-up procedure by taking artificial agents equipped with cameras and presenting them with a number of photographs, which served as a base for him to teach them some basic two-dimensional spatial relations – the experiment was not entirely unsuccessful, but it seems apparent that it achieved only basic machine learning rather than grounding.

The approaches above are both cognitivistic; that is to say they belong in the group of approaches to various mind-related problems that distance themselves from agents' behaviours and rather focus on underlying processes within the mind that elicit said behaviours. However, neither of them yielded desired results, which some saw as bad news for cognitivism in symbol grounding in general. Ziemke (1998)

argues that this is because they are simply tagging things out there with prescribed symbols, and do not interact with them enough to be able to achieve grounding. Ziemke therefore proposes what he calls enactive grounding. This approach is based on true interaction between the artificial agent and its environment, which calls for agent embodiment, i.e., the agent must be given a physical form that allows it to interact as much as possible, therefore including visual, audio, and any other possible receptors. With such a system, it is also possible to arrive at behaviour emergence (a behaviour emergent from agent's interactions, independent from its source code or such), and, by extension, grounding emergence. Another example of an enactive approach is Sun's (2000) approach, which mainly relies on phenomenology, claiming that an agent has to be embodied to be in the world and to be able to make itself available for recognition in the world. Both of these enactive approaches are facing the externalist trap, which is the reduction of agent's behaviour to mere reactions to external factors in its environment, placing the environment first. If all behaviour of the agent is nothing but a reaction to outside stimuli, then the agent cannot be considered autonomous (Ziemke 1998). This is part of a more fundamental question of how exactly an artificial agent and the environment would interact, beyond the AI simply recording the environment and again, merely tagging things with symbols.

Enactivism has generally proven to be a rather popular approach within cognitive science and can be primarily described as a position that seeks to explain cognition and mental processes as a complex set of interactions between a living agent, its immediate environment, and the world in general (Varela, Thompson and Rosch 1991). According to these authors, enaction itself is the process in which a perceiving agent acts (either consciously or automatically) to the requirements of its environment and given situation. This basic form of enactivism is known as autopoietic enactivism, where autopoiesis refers to the process of self-maintenance and autonomy It is supposed to both present an alternative to dualism in the sense that the distinction between mental and biological processes is almost eliminated, and the former seem to supervene on the latter, as well as distance itself from representationalism (Maturana and Varela 1992).

The notion of this distancing is better shown within the theory of sensorimotor enactivism, which claims that perception is an active, rather than passive process, where perceiving agents actively explore and intentionally seek to interact with the world. In those interactions, they appeal to sensorimotor expectations about how objects in the world will change depending on the agent's angle of perception, physical interactions with said objects, etc. (Noe 2004). These expectations are what then defines cognition, and are considered to be non-representational, although it could be argued that they still demand some degree of mental modelling of the expected states of the world.

Finally, theories of radical enactivism seek to eliminate representation altogether. Hutto and Myin (2013) for example go to great lengths to deconstruct various preceding views of cognition, including those found in autopoietic and sensorimotor enactivism, in order to explain them purely in terms of enaction and without any need for representation. Surprisingly, they arrive at the conclusion that representations can be avoided only on the level of basic cognitive and perceptual processes, i.e., when dealing with concrete objects and concepts, and that complex processes such as language nevertheless need to rely on representations to process abstractions and symbols in language.

These enactive approaches are therefore all still based in representationalism. Although they seek to distance themselves from representationalism, autopoietic approaches never claim they have done so entirely, sensorimotor approaches seem to revert to them at least partially when one considers how exactly "expectations" are manifested in the agent, and radical approach admits it is only possible on rather basic levels. Theories of enacted cognition have great potential in pursuit of grounding in artificial agents as they complement embodied cognition remarkably well, as well as present an adequate basis for (symbol) emergence, which we will mention later. Now, however, we shall return to our analysis of other various approaches.

Next to be explored is the functional model developed by Mayo (2003), where what Harnad considers categorical representations are interpreted in a functionalist sense. Every category is considered a set that contains functionally relevant elements. A single symbol may evidently therefore exist in several functional categories. Mayo claims that it is this very overlap in functions of one discrete symbol that characterises it as distinct from those who share some but not all its functions. These various representationalist approaches are important because newer approaches to the symbol grounding problem tend to return to representationalism at least in the early stages of the grounding procedure. Still, we shall briefly mention semi-representationalist and non-representationalist approaches as well.

One of the semi-representationalist is, for example, the physical symbol grounding problem, where a symbol is considered a physical form of what is represented. A semiotic symbol system consisting of form, meaning, and referent, is introduced; in that, the form is the physical tag of a symbol, the meaning the semantic content of the physical tag, and the referent is the "thing" in the outside world to which the tag applies. Artificial agents then attempt grounding through an imitation game consisting of speaker agents and hearer agents. The speaker agents vocally express the symbolic tag of the referent, while the hearer agents must figure out what it applies to. The idea is that the symbol (symbolic tag) is grounded in the hearer agent when it can accurately recognise the referent upon hearing the tag (without intermittent mistakes) (Vogt 2002). Finally, non-representationalist mod-

els entirely disregard representations' role in the symbol grounding problem and instead fully rely on the interaction between the artificial agent and its environment.

A breakthrough is made by Taddeo and Floridi (2005), who review all significant research on the topic since Harnad, finding that none of the approaches above, as well as numerous others we left out in this analysis, satisfies what they call the Zero semantic commitment condition or Z-condition for short. The latter is formalised as follows:

> 1) No form of *innatism* is allowed; no semantic resources (some *virtus semantica*) should be presupposed as already pre-installed in the AA; and
> 2) no form of *externalism* is allowed either; no semantic resources should be uploaded form the "outside" by some *deus ex machina* already semantically-proficient.
> Of course, points (a)-(b) do not exclude the possibility that
> 3) the AA should have its own capacities and resources (e.g., computational, syntactical, procedural, perceptual, educational etc., exploited through algorithms, sensors, actuators etc.) to be able to ground its symbols. (Taddeo and Floridi 2005: 423)

Most forms of approaches we described above rely on innatisms, which are indeed problematic for symbol grounding, but some merely rely on certain externalisms, that we will later argue can be sometimes justified in analogy to human grounding.

The same authors later (2007) establish their approach to symbol grounding that they claim satisfies the Z-condition and brings one as close as possible to solving the problem in question. The first principle they introduce is the Action-Based Semantics, which assumes that meanings are in their first stage internal states of the agent, whereafter they trigger actions, which proves them to cause semantic emergence in the agent (without innatism). The second principle is the division of the agent into two machines that both communicate with the environment and each other, thereby allowing the agent to reflect on its actions. This latter principle allows access to communication capacities, categorisation/abstraction capacities, and representational capacities within the agent, as well as access to feedback. The former principle provides a sensomotorical interactive approach, as well as an evolutionary approach and the satisfaction of Z-condition.

As successful as this approach may seem, Bringsjord (2014) emphasises that Taddeo and Floridi's approach lacks the explanatory power as to how an artificial agent, functioning based on their design, could reach the level of grounding where it could communicate on the same level as a competent human speaker. Bringsjord invokes an example of a letter written by a girl to her boyfriend, which a human reader (such as me or you) can plainly understand to be sarcasm; a good approach to grounding must be able to explain how an artificial agent can reach the level of understanding sarcasm, humour, pragmatics, metaphors, etc. Bringsjord himself notices that Z-condition might be blocking that

possibility entirely on higher levels of grounding, while evolutionary approach to grounding also seems to be quite faulty.

The issue with the evolutionary approach is that there is no concrete evidence that human linguistic competence developed strictly through evolution since some early linguistic features were quite redundant as per humans' needs at the time (Bringsjord 2014); it is also hard to grasp how simulating the entirety of human language evolution in an individual artificial agent would make any sense. As Harnad (1990) implies in the Chinese Merry-go-round description, an artificial intelligence attempting symbol grounding is not unlike a baby learning its first language, and by no means does a baby lying in her crib have to invent words for things she sees around here. She will not replicate linguistic evolution and emerge at 18 months old with a private language, rather, she will learn the language(s) of her parents by interacting with them and their environment, and it is likely this principle of human language development we should follow when pursuing symbol grounding.

### 2.2. *Human grounding simulation*

The first thing that seems to be quite on point about this notion is that it is evident that children learn their first language – for which they have to acquire symbol grounding – through interaction with their environment (Vogt 2007). The children learn their first language by attributing meanings to symbols depending on the symbols' context in terms of both other symbols as well as perception data available (e.g., if someone is pointing at a particular thing when uttering a symbol). The agent must decide on a symbol's meaning depending on all of its contextual features. Vogt serves an example where a linguist hears a native speaker of an unknown language utter "Gavagai" when a rabbit appears on the scene. Purely logically, the auditory symbol "gavagai" could mean numerous things, but for humans it is intuitively very easy to determine its most likely meaning is "rabbit." We may remark here that the original use of the Gavagai example appears in Quine (1960), where the linguist in question undergoes a tedious procedure of verifying her assumption that "gavagai" is more likely to mean "rabbit" than "white" or merely "animal" by studying the natives' affirmative and negative responses to her using "gavagai" in those varying contexts. Our point here, however, relates to none of these. Rather, what we wish to take away from this example is how easy it is for humans to intuitively grasp the most likely meaning of a new word, immediately favouring the more likely "rabbit" over less likely but plausible "white" or "animal."

An artificial agent, however, may have trouble recognising instances on its own, therefore it would likely require some prerequisite competencies that would allow it to be able to make such a connection as the one between "gavagai" and a rabbit. It should, for instance, somehow

know what it means when somebody points at something, as humans seem to intuitively even at a very young age. We are, again, not claiming that humans can determine with utmost certainty the meaning of any new word; we are simply observing that we seem to have a predisposition to pick out the most likely of various possible meanings with a decent degree of success.

Cowley (2007) offers a solution to this dilemma when he describes that a (human) baby primarily relies on the role of its parents when learning to communicate. Namely, it relies on the notion that its parents will demonstrate, by communicating to it and each other, an appropriate pattern of actions, vocalisations, and relations between action and vocalisation. What happens in this procedure is that children learn to speak by being explained or shown symbols their parents have already grounded. Children finally become competent speakers by coordinating with the others consistently in a certain cultural or social environment. In reference back to Quine and Vogt's Gavagai example, a child gets to know the meaning of "rabbit" from being shown a rabbit (or an image thereof) by her parents, accompanied by them uttering the word "rabbit," presuming they know what a rabbit is and that the symbol "rabbit" refers to that particular fluffy creature. In a later circumstance, the same child, now adult, will assume (likely correctly) that "gavagai" means "rabbit" rather than "white", because that is how her parents demonstrated new symbols. Of course, in this later context, Quine's verification procedure applies, as it does for artificial agents, which we will show later, noting also that for artificial agents all possible meanings of a symbol carry the same probability value, which is not true for human agents. For Cowley, there is also no pure symbol as far as humans are concerned – rather, symbols are a posteriori and derived from the use of language, grounded in behaviour and action.

Another type of simulation that we may require to achieve grounded cognition is a more direct simulation of cognition itself (Barsalou 1999, 2008; Pezzulo et al. 2013). Barsalou (1999) proposes an approach named Perceptual Symbol Systems theory, which acknowledges that modal symbolic operations are of great importance for interpreting experience and suggests that natural implementation of such operations can be achieved by the means of mental simulations. According to the theory in question, there is "a single, multimodal representation system in the brain that supports diverse forms of simulation across different cognitive processes" (Barsalou 2008). Such cognitive processes include several types of perception, various levels of memory, as well as conceptual knowledge. This allows for (multimodal) states to be captured in memory and retrieved to be simulated when required. These processes occur within human cognition, as well as, according to Barsalou, in non-human agents (in this case, animals). Reasonable assumption is that such systems of mental simulation should also be computationally emulated within artificial agents to achieve grounded cognition and in turn symbol grounding (Pezzulo et al. 2013).

Barsalou (2008) emphasises on the link between language and simulation, pointing out several examples: situation models, perceptual simulation, motor simulation, affective simulation, and gestures. Situation models are spatial representations, or better yet, spatial situation simulations that occur when scenes from written texts are described to people verbally, showing a tight relation between visual and verbal comprehension of spatial situations. Perceptual simulations refer to the representations an agent constructs when a concrete object is described to them; when a description of an object is vague, the representations contain implicit perceptual information about the object, which is more than likely drawn from the agent's memory. Next, motor simulations occur when verbs for actions of various body parts are described to the agent, which triggers a reaction in their motor system; according to Barsalou (2008) neurological research had shown this happens on the level of the central nervous system even when the corresponding action does not manifest physically. Fourth, affective simulations are those that occur when an agent is exposed to a word, or a text, that carries some form of emotional value for the agent. Finally, gestures are an expression of embodiment in language that connect bodily movements with the meanings of words they accompany. Barsalou (2008) provides numerous examples from empirical studies that support all of the above types of simulation-language relations. However, such examples are hardly in the scope of this paper, but we encourage the reader to refer to the original text by Barsalou.

Grounded cognition through mental simulations as summarised above can greatly contribute to achieving symbol grounding; a great additional illustration of this can be found in Pezzulo et al. (2013) where the authors explain the "cascade of effects on cognition" from grounding through embodiment to situatedness. It also concurs with the requirement for human-grounding simulation we have discussed at the beginning of this section (in Vogt 2007 and Cowley 2007), as well as with requirements for multimodality and embodiment (e.g. Ziemke 1998; Cangelosi and Riga 2006).

We would like to pause to address the issue we mentioned with Taddeo and Floridi's Z-condition. Particularly that the second point of their condition, which prohibits any and all kinds of externalism is too stringent. If we look back to Cowley, we see that children seem to learn at least in part by being explained symbols by agents who are already semantically proficient, that is to say they have already grounded those symbols. A simple example of this would possibly be a child's mother pointing at herself and saying "momma" when interacting with her toddler. Eventually, every healthy child will successfully learn that "momma" is that female figure that feeds her, consoles her, plays with her, etc., and learn to point at her and say "momma" as well. At later stages, the child may be attending school, where she is very plainly explained the meaning of the word "addition" in mathematics or "gravity" in physics. If such externalist explanations do not violate human

grounding process, why should they be considered as violations of artificial agents' grounding processes? Indeed, without such externalism we seem to be forever stuck on a version Harnad's impossible version of the Chinese Merry-go-round where we expect an agent to learn a first language from a dictionary. To prevent such conundrums, certain externalisms have to be allowed in the grounding process. Of course, the process should not be fully reliant on them, as children learn plenty by simply observing what others vocalise in different contexts and learn to replicate that quite successfully on their own.

This can greatly contribute to what is already known in robotics as the epigenetic model and can feature in Emergent symbol grounding approaches (Tangiuchi et al., 2016). The latter proposes that in humans, symbols emerge throughout the language learning process, wherein they automatically connect to referents and each other, thereby grounding themselves in perceptions, internal representations of those perceptions, and actions. Tangiuchi et al. introduce their own requirements for this model to be successful. One of those is, for instance, multimodal categorisation, which requires agents to ground every category (of things) in multiple modalities, i.e., visual perception, audio perception, haptic perception, and any others available. Thus grounded (categorical) symbol includes all perceivable features of the thing or all common perceivable features of the category of things in which it is grounded.

An interesting example of an early epigenetic model is Cangelosi and Riga's (2006) experimental embodied agent. They suppose two grounding mechanisms: the first grounds basic vocabulary directly in environmental interaction; the second one is transferred grounding that allows the agent to join two basic grounded elements and ground in them a more complex symbol. We will not go into many details of the experiment. The robots had a number sensomotoric actions in their programming but lacked any symbol to connect them with – upon receiving a symbolic order, such as "Close left arm," they randomly performed one of those actions and received positive feedback if right. The first phase consisted of repeating this procedure on several basic phrases. The second phase contained phrases such as "Grab" and the agents had to "figure out" that "Grab" consists of "Close left arm and Close right arm." In the third phase, they had to ground phrases that were conjunctions of the second phase phrases. The experiment was rather successful with a high rate of accuracy on all three stages; however, even the basic stage required a large number of repetitions, with the second and third requiring respectively more. This can be partially ascribed to the processing power of computers fifteen years ago, or we can say that perhaps symbol grounding is a procedure that is just as long and complex as first language learning is in children.

What have we ended up with at this point? It seems like that in order to achieve grounding, we require:

1. An embodied agent with multimodal capacity

2. An epigenetic approach to symbol grounding (simulating human first language acquisition and human cognition in terms of mental simulations)
3. A multi-phased approach to symbol grounding (allowing complex symbols to be grounded in baser symbols, or to be simply explained)
4. For the purposes of 2 and 3: dropping the second requirement of the Z-condition
5. To be prepared the procedure may take a very long time (as a consequence of 2)
6. Explanatory Power for the Non-stupidity Condition

It is this last point the second half of our article will focus on.

## 3. *Explanatory power for pragmatics*

If we are to move on to satisfying the Non-stupidity condition, the first thing we ought to do is explain how grounding abstract symbols can be achieved as some nth phase of our multi-phase grounding model, wherein the early phases involve grounding very concrete, physical symbols with increasing complexity. What we consider an abstract symbol is a symbol without a physical or directly perceivable (by means of multimodal sensory apparatus) referent in the outside world (Cangelosi and Riga 2006; Šetar 2020b; Tangiuchi et al. 2019)

Initially, some basic symbol grounding is quite correctly described already by Harnad, albeit in a representationalist way. Harnad claims that once we have grounded both the symbol "horse" and the symbol "stripes" – in this case we are grounding them nicely and slowly through epigenetic, multimodal interaction – we can ground the term "zebra" without actually having any experience with the primary referent for "zebra." It is enough that an agent experiences pictures or films of a zebra but can also form an idea of a zebra as a black-and-white striped horse similarly as "horn" and "horse" can lead to the idea of a unicorn. However, these sorts of conjunctions only seem to function as far as concrete symbols with physical referents are concerned.

To understand how grounding might proceed for abstract concepts and pragmatic elements, we can look at four requirements proposed by Tangiuchi et al. (2019):

> – Creating holistic language processing systems that involve physical, psychological, social, conceptual, and experiential constraints.
>
> – Inventing machine learning methods to represent the recursive property of background beliefs for holistic language processing.
>
> – Developing computational models for collaborative tasks in the physical world, leading to the emergence of dialogue.
>
> – Inventing methods to enable a robot to make use of contexts, e.g., situation and culture, and to grow the ability to use language to exchange meaning by referring to social factors: field, tenor, and mode. (Tangiuchi et al. 2019: 20)

While developing computational models is a matter best addressed by those with more technological prowess than the authors of this article, and inventing machine learning methods, even just theoretically, is a detailed and tedious task that falls out of the scope of this article, the first, and especially the latter requirement may shed some light on the issues at hand. Tangiuchi et al. (2019) look for a solution in Halliday's functional linguistics, where the semantics of a word depends on its contextual use, depending on culture and particular situation. And while situational and cultural contexts may be taught to artificial agents with some additional effort, we shall seek a solution elsewhere – namely, Chomsky's theory of universal grammar (1957). The idea we are focusing on here is that every sentence has a kernel unit, while the sentence is a transformation of that kernel. The transformation itself is not a matter of semantics but rather a tool for the disambiguation of meaning based on socially defined functional semantics. This offers us an option to pre-equip our artificial agent with a non-semantic grammatical apparatus that enables syntactic formation and transformations; the latter are defined by the interaction of our agent with its environment, which teaches it, by providing examples to be analysed, which transformation is correct in what context. The sentence kernels are those symbols that need to be grounded in the traditional sense. Additional insight is offered in more recent Chomsky (2017), where the author determines that given the speed at which language is acquired by children and the low amount of presentation required for them to learn and ground a new linguistic symbol, language itself or at least the basis thereof must be deeply internalistic and supervene on simple computational processes, with all externalisms coming in later, allowing for communicative faculties of language. While some other aspects of the article in question pose some new issues for language grounding in artificial intelligence, mainly in the environmental interactivity department, there is an important new point to be made. If language is, when sufficiently reduced to its evolutionary core, indeed a simple computational process, then this computational process may be quite easily replicated in artificial neural networks once it is determined how it works on a formal computational level in humans. The notion that the (generative) acquisition of one's first language is deeply internalistic and requires very few presentations, also entails that the internalist trap (the opposite of the externalist trap defined earlier) is not in fact a trap, but a necessary first step in language development. Referring to Vogt and Quine's example, we learn the word "rabbit" via a computational, internalist process that pertains to acquiring one's first language, and later affirm it and attempt to disambiguate "gavagai" in virtue of second-order externalist processes that pertain to effective communicational use of our first language as well as acquiring further languages. Much further ado is necessary here, which would only confuse the rest of this article, but may serve as a basis for an entirely separate one in the future.

Going back to allowed preconditions for symbol grounding – field, tenor, and mode: all of the elements are in and of themselves non-semantic and could therefore be used as a tool in an epigenetic model for symbol grounding. However, the field requires an understanding of topics, and cultural and social context, which can only be learned through interaction and communication; therefore, it cannot be precluded in an agent. We have similar issues with mode, which characterises discourse structure, way of expression, etc; again, slangs, registers and such must be learned as part of satisfying the Non-stupidity Condition. Lastly, however, some parts of tenor may be precluded in a learning agent. While it will develop social relations with other agents on its own, it is in no contradiction with epigenetic modelling to pre-equip an artificial agent with devices that allow it to perceive certain tones of voice, pitches, etc. as negative or positive, seen as a baby has no issue distinguishing between, for example, a parent being upset and a parent being caring.

Another concept that may be required to proceed from concrete symbol to abstract symbol grounding is the concept of semantic affordance (Glenberg and Robertson 2000). A chair, which can basically be defined as a piece of furniture with four legs affords humans with a function of sitting but does not afford the same function to an elephant, while it affords this function to a cat only incidentally but not intentionally. There are also contingent affordances, such as the affording the function of being stood on to reach a higher location.

It is multimodal sensory experience that first helps ground the notion of "chair" and it also helps extend this notion to a variety of chairs – those with three legs, those without a back, etc. It is at a later stage that "leg [of a chair]" is grounded as part of a chair and distinctly from "leg [of a human]." However, "chair" is a very simple, concrete symbol, and so is "leg [of a chair]," even though it is located a phase higher in grounding hierarchy than "chair."

Finally, let us look at how one could ground "[a] painting." In the earliest multimodal grounding phase, we would need an experience of seeing a number of depictions of things, which are not photographs and not printed in any other form; haptic perception (i.e., touch) could be of help here in recognising the texture of a painting. Next, we would need to have already grounded concepts of "form [in general]" and "content [in general]," which an agent would then have to specialise to "form [in painting]" and "content [of a painting]" – this can be done by explaining the agent how these concepts work in painting just as an art teacher would explain it to students. Several stages later, a complex grounded scheme like "(if 'form' is 'dynamic'… and 'content' is 'exaggerated,' 'twisted'…)" can mean "expressionism." These notions are extremely difficult to describe in humans, not to mention in artificial agents. The point is, however, that in humans such multi-layered approach to grounding evermore complex and abstract symbols seems to work – therefore, why should it not in a sophisticated epigenetic artificial agent?

### 3.1. *In speech acts*

While speech acts were first formulated by Austin (1962), we will not use his threefold classification (locution, illocution and perlocution) in our attempt to describe possible grounding mechanism for speech acts because we argue (Šetar 2020a and b) that locution, illocution and perlocution are in fact features of speech acts that every speech act possesses.

Instead, we will use a more contemporary classification of speech acts into the assertive, commissive, constative, directive, and imperative speech acts (Jary 2010; Kissine 2013; Jary and Kissine 2014). Assertive speech acts are statements that are truth-bearing and convey truth-value information without explicit intention of altering the hearer's belief; commissive speech acts are ones that speaker uses to commit themselves to fulfil their content, such as promises and threats; constative speech acts are ones intended to alter the hearer's belief regardless of their de facto truth value; directive speech acts intend to convince the hearer to fulfil their content by providing sufficient reason to do so; lastly, imperative speech acts instruct the hearer to fulfil their content without providing a reason but rather do so by other means, most commonly by being uttered from a position of authority. The five classes of speech acts can be formalised as follows:

> Assertive: "A is and assertive speech act containing proposition p if, and only if, the speaker believes p to be true and there is justification for p to be true." (Šetar 2020a: 35, drawing on Jary 2010)
> Commissive: "All promises are acts of placing oneself under an obligation to bring about the propositional content p." (Kissine 2013: 149)
> Constative: "An utterance is a constative speech act with the content p if, and only if, with respect to this background, it constitutes a reason to believe that p." (Kissine 2013: 62)
> Directive: "An utterance is a directive speech act with the content p if, and only if, with respect to a given background, it constitutes a reason to bring about the propositional content of p." (Šetar 2020a: 44, drawing on Kissine 2013)
> Imperative: "I is an imperative speech act containing proposition p if, and only if, it compels the hearer to bring about the propositional content of p." (Šetar 2020a: 46, drawing on Jary and Kissine 2014)

But why do we require such formalisations in the first place? That is due to the fact that humans recognise the function and intention of speech acts entirely intuitively, that when hearing a certain phrase, we do not have to break it down and consciously consider what speech act it is, we simply know. This could be an inherent faculty of ours being conscious, and since it would be terribly reductive for one to assume that symbol grounding or any other form of artificial intelligence entails consciousness (see Pierce 2017), we must find a mechanism to teach speech acts to an agent that is not necessarily conscious and does not necessarily possess intuitions or other such capabilities. Given the logical nature of programming and computer operations, logical formalisations of speech acts are a reasonable way out. However, we need

a concrete symbolic referent through which a speech act can be determined to belong to a certain class. In Šetar 2020a we found that a viable candidate for this in English may be modal verbs, which can also be nicely logically formalised:

> Can: p is compatible with the set of all propositions which have a bearing on p.
> May: there is at least some set of propositions such that p is compatible with it.
> Must: p is entailed by the set of all propositions which have a bearing on p.
> Should: there is at least some set of propositions such that p is entailed by it. (Where p is the proposition expressed by the rest of the utterance). (Papafragou 1998: 50)

Modals "have to" and "ought" to be also considered here; for the purposes of this article "have to" is seen as an equivalent of "must", and "ought" is formalised between "must" and "should," as it is generally perceived as deontically weaker than "must", yet stronger than "should". We explain this in a bit more detail in Šetar (2020a), where we draw on Groefsema's (1995)'s formalisations of modals "must" and "should," also summarised in Papafragou (1998). If in "must," the contained proposition p is entailed by all prepositions that have a bearing on it, and in "should" it only needs to be entailed by some arbitrarily small set of such propositions, we can say that in "ought," p is entailed by most of the propositions which have a bearing on p.

What this brings us is the notion that assertive speech acts can be those that are either non-modalized or involve entailing modals "have to," "must" and "can" in an epistemic sense, which is to say they convey a certain knowledge or belief. The need for strong entailing modals arises from the fact that assertive speech acts necessarily convey knowledge and not mere belief.

Unlike assertive speech acts, constative speech acts are ones intended to convince the hearer of speaker's belief (not necessarily knowledge), they can feature any modal used in an epistemic sense. For example, "there should be a connection between those events" is a constative speech act, and so is "they must be brothers." However, "increasing summer temperatures must be related to global climate change" is an assertive speech act.

For commissive speech acts we can say they are those using "must" and "have to", as well as sometimes "ought to" in first person, in a deontic way – the latter meaning that they express a duty to do something: specifically, to bring about the proposition contained in the utterance. "Will" can also be considered a modal verb that shows intention to do something and can therefore also be an indicator of a commissive speech act.

Directive speech acts are also based on deontic use of modals and, like assertive speech acts, require entailing modals, albeit not only the stronger ones. Thus, "you must finish your homework" and "you should not be late again" are both directive speech acts. They can, however,

also be imperative, depending on what kind of deontic justification lies behind their use. If the former is spoken by a teacher and the latter by a boss, they are justified by authority and therefore certainly imperative – yet if they are uttered by the hearer's friend they are directive, as they are otherwise justified, for example as "you must finish your homework [if you wish to pass the course]" and "you should not be later again [if you wish to avoid disciplinary action]".

It is reasonable to also mention performative speech acts, which are difficult to formalise in the way presented above, as they are speech acts that alter something in social (conventional) reality, if uttered from a position of proper authority. Notable examples are "I now pronounce you man and wife" as uttered by a priest, or a parent naming their new-born child.

Even though a modal verb can be an excellent cue for the artificial agent to start identifying a speech act as belonging to a certain class and having a certain function, it does not fully define a speech act. What is still necessary is for the artificial agent to have certain conception of epistemic and deontic use, as well as of authority. This is where we refer back to the emulation of grounding development in humans and pre-given capabilities related to recognising tone, mode, and field of discourse. An artificial agent with a long enough learning process will have grounded the concept of authority relatively early in that process and will be able to distinguish different uses of the same modal verb depending on the pattern of their use by others. That is to say that it should be able to conceive of "you must clean this room" as imperative or directive based on the deontic "must", while also being able to understand that "you must try these cookies" is in no way an imperative or even a directive, based its interaction with environment, i.e. based on how "must" is usually used, in what contexts it is used, and how human hearers react to it depending on its uses in different contexts.

### 3.2. *In metaphors*

An important aspect of satisfying the condition of non-stupidity is accounting for how metaphorical speech may be grounded since that very type of speech is commonplace in everyday communication in idioms, proverbs, literature, etc. In doing so we will first refer to the notion that metaphorical utterances can be understood in two ways: through their original domain or through the target domain (Tangiuchi et al. 2019). The original domain involves concrete symbols and concepts whose referents are usually empirically accessible, i.e., the literal meaning of the phrase, while the target domain is the translation of those symbols and concepts into their abstract meaning, which is semantically related to the literal meanings in the original domain.

Let us examine the idiom "she wouldn't harm a fly." If this idiom was to be understood in context of its original domain it would be in-

terpreted as if the person in question has an actual, literal aversion towards harming a particular type of insect. That sort of interpretation is certainly stupid in Bringsjord's sense. In context of its target domain, however, it means that the person to whom the metaphor refers is very peaceful and gentle. Where that derives from is the conception that striking a buzzing fly is generally considered an extremely mild, or even the mildest conceivable form of violence. To say that someone is not willing to cause (even) that much violence is to say that they would certainly not commit any act more violent than that, therefore that they would not commit any act of violence at all.

The semantic link between the original and target domain implies that every metaphor can be broken up into non-abstract elements, therefore the primary condition for being able to ground and understand metaphorical expressions is to have already grounded the necessary non-abstract symbols, which we optimistically claim may be well achievable through the methods we described earlier.

For the second step, we need to know how an artificial agent may be able to understand the translation of original domain into the target domain. In humans we can claim this happens through being exposed to idioms and such simple metaphorical expressions in their interaction with others, which, if the embodied symbol-emergence based approach we have been advocating for holds, is likely to happen in any learning artificial agents with proper grounding capabilities described at the end of section 2.2. Here, it is also worth noting that some extremely common idioms, such as the one used in our example above, may also work the other way around: an agent, human or artificial, commonly exposed to the use of this particular idiom, may, for example, learn that harming a fly is the lowest form of violence through being exposed to the metaphor.

Another approach that may yet better coincide with our requirements for embodiment and human cognition simulation is found in the works of Lakoff and Johnson (1980, 1999, 2003), namely in their notion of a conceptual metaphor. The latter argues that metaphors do not pertain only to language but to cognition in general. That is to say that humans tend to utilise metaphors not only to express themselves but also to think about things on conscious and unconscious levels. The latter concept of unconscious processing of metaphors is called functional embodiment and observes that certain concepts, including conceptual metaphors, are used automatically in cognitive processes without conscious awareness of the agent, as opposed to only being understood on an intellectual level (Lakoff 1987). This leads to some interesting implications about metaphorical mapping (i.e. the mental transition from the source domain to target domain, as well as translation from target to source) as a subconscious cognitive tool used automatically to process and describe perceptions and experience, as well as to interpret verbal inputs in metaphoric form, which may give rise to the category

of metaphorical simulations, which we can fit in with Barsalou's (2008) mental simulation categories (our thanks to one of the anonymous reviewers for pointing this out). Lakoff and Johnson (1980/2003) somewhat controversially go as far as to say that metaphor mapping may be directly related to the way our brains are mapped – this, if true, practically guarantees that if proper grounding is achieved as we have described in section 2, conceptual metaphor mapping will emerge in an embodied, interactive agent.

Lastly, we may also conceive of how literary metaphors may be grounded – through exposure to common idioms, an agent learns what metaphorical meanings certain symbols commonly hold, for example that fire is often metaphorical of life, or flame of passion, etc. The process is completely analogous to one of function affordance by Glenberg and Robertson (2000) that we have described earlier. Further, there are certain metaphors in literature that are entirely unique and their meaning is speculated about by literary analysts – in these cases it is perfectly acceptable for an artificial agent to have ability of exercising such speculations, making non-stupid guesses based on its previous experience of metaphors, as we do not expect it to possess a magical insight into the mind of the metaphor's creator. However, this does not need be the case; an important part of Lakoff and Johnson's idea of conceptual metaphors is that metaphors may be grounded in simpler metaphors (equivalences, such as "love is war") that can then produce a virtual infinity of related metaphors (see also Pinker, 2007), and are themselves grounded in concrete concepts that are perceptually and experientially accessible and then serve as source domains to be related and mapped into target domains when metaphors are formed or analysed.

## 4. *Conclusions*

What we ultimately provided here is a theoretical approach to symbol grounding that merges compatible elements of prior prominent models of symbol grounding, including embodied agents, long-term learning that emulates human first language learning process, and symbol emergence theory, which has the explanatory power with which it can satisfy Bringsjord's (2014) non-stupidity condition.

The explanatory power lies in being exposed to a vast amount of language symbols through interaction with the environment over a long period of time, through which process an artificial agent builds a database of various contextual uses of individual symbols and from it learns to correctly determine the meaning of a symbol in certain context – a process which allows for grounding of specific contextual affordances of symbols, such as metaphoric ones, and predicting (guessing) the meaning of symbols in first-time-seen contexts.

Despite being quite successful at explaining these already high-order levels of grounding, the approach has certain limitations. For example, it remains to be determined, how certain elements of human

communication, such as sarcasm, irony, or humour could be understood or grounded by artificial intelligence, even though we have hinted that the solution may lie in pre-given capabilities related to identifying tone of discourse and similar elements. Therefore we have approached satisfying the non-stupidity condition, but there are still certain questions to be answered before the explanatory power of this working model is entirely adequate.

## *References*

Barsalou, L. W. 1999. "Perceptual symbol systems." *Behavioural Brain Science* 22: 577–660.

Barsalou, L. W. 2008. "Grounded Cognition." *Annual Review of Psychology* 59: 617–645.

Bringsjord, S. 2014. "The symbol grounding problem ... remains unsolved." *Journal of Experimental & Theoretical Artificial Intelligence* 27 (1): 63–72.

Cangelosi, A. and Riga, T. 2006. An Embodied Model for Sensorimotor Grounding and Grounding Transfer: Experiments with Epigenetic Robots." *Cognitive Science* 30: 673–689.

Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. 2017. "The Galilean Challenge: Architecture and Evolution of Language." *J. Phys.: Conf. Ser.* 880 012015.

Cowley, S. J. 2007. "How human infants deal with symbol grounding." *Interaction Studies* 8 (1): 83–104.

Davidsson, P. 1993. "Toward a General Solution to the Symbol Grounding Problem: Combining Machine Learning and Computer Vision." In *AAAI Fall Symposium Series, Machine Learning in Computer Vision: What, Why and How?*, 157–161.

Fodor, J. A. 1980. "Methodological solipsism considered as a research strategy in cognitive psychology." *Behavioral and Brain Sciences* 3: 63–69.

Glenberg, A. M. in Robertson, D. A. 2000. "Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning." *Journal of Memory and Language* 43: 379–401.

Groefsema, M. 1995. "Can, may, must and should: A relevance theoretic account." *Journal of Linguistics* 31: 53–79.

Guazzini, J. 2017. "An Epistemological Approach to the Symbol Grounding Problem." In V. C. Müller (ed.). *Philosophy and Theory of Artificial Intelligence.* Leeds: Springer, 36–39.

Harnad, S. 1990. "The symbol grounding problem." *Physica D* 42: 335–346.

Hutto, D. D. and Myin, E. 2013. *Radicalizing Enactivism: Basic Minds without Content.* Cambridge: MIT Press.

Jary, M. 2010. *Assertion.* London: Palgrave Macmillan.

Jary, M. in Kissine, M. 2014. *Imperatives.* Cambridge: Cambridge University Press.

Kissine, M. 2013. *From Utterances to Speech Acts.* Cambridge: Cambridge University Press.

Lakoff, G. 1987. *Women, Fire, and Dangerous Things.* Chicago: The University of Chicago Press.

Lakoff, G. and Johnson, M. 1980/2003. *Metaphors We Live By. With Afterword 2003.* Chicago: The University of Chicago Press.

Lakoff, G. and Johnson, M. 1999. *The Embodied Mind and Its Challenge to the Western Thought.* New York: Basic Books.

Maturana, H. R. and Varela, F. J. 1992. *The tree of knowledge: the biological roots of human understanding.* Boulder: Shambhala Publications.

Mayo, M. 2003. "Symbol Grounding and its Implication for Artificial Intelligence." *Twenty-Sixth Australian Computer Science Conference*, 55–60.

Müller, V. C. 2015. "Which Grounding Problem Should We Try to Solve?" *Journal of Experimental & Theoretical Artificial Intelligence* 27 (1): 73–78.

Noe, A. 2004. *Action and Perception.* Cambridge: MIT Press.

Papafragou, A. 1998. *Modality and the Semantics-Pragmatics Interface.* London: University College London.

Pezzulo, G. et al. 2013. "Computational Grounded Cognition: a new alliance between grounded cognition and computational modelling." *Frontiers in Psychology* 3: 1–11.

Pierce, B. 2017. "How Are Robots' Reasons for Action Grounded?" In V. C. Müller (ed.). *Philosophy and Theory of Artificial Intelligence.* Leeds: Springer, 73–80.

Pinker, S. 2007. *The Stuff of Thought.* London: Penguin Publishing Group.

Quine, W. V. O. 1960. *Word and Object.* Cambridge: MIT Press.

Regier, T. 1992. *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization.* Berkeley: Department of Computer Science, University of California at Berkeley.

Rodriguez, D. et al. 2011. "Meaning in Artificial Agents: The Symbol Grounding Problem Revisited." *Minds and Machines* 22 (1): 25–34.

Searle, J. R. 1980. "Minds, brains and programs." *Behavioral and Brain Sciences* 3: 417–457.

Shepard, R. N. and Cooper, L. A. 1982. *Mental images and their transformations.* Cambridge: MIT Press/Bradford.

Steels, L. 2008. "The symbol grounding problem has been solved, so what's next?" In M. de Vega. et al. (eds.). *Symbols and embodiment: Debates on meaning and cognition.* Oxford: Oxford University Press, 223–244.

Steels, L. in Vogt, P. 1997. "Grounding adaptive language games in robotic agents." In C. Husbands and I. Harvey (eds.). *Proceedings of the 4th European Conference on Artificial Life.* Cambridge: MIT Press.

Šetar, N. 2020a. *A Monosemic Account of Modality in Speech Act Theory. MA Thesis.* Maribor: Univerza v Mariboru.

Šetar, N. 2020b. *Utemeljevanje simbolov in pragmatika v umetni inteligenci. MA Thesis.* Maribor: Univerza v Mariboru.

Taddeo, M. in Floridi, L. 2005. "Solving the symbol grounding problem: A critical review of fifteen years of research." *Journal of Experimental and Theoretical Artificial Intelligence* 17 (4): 419–445.

Taddeo, M. in Floridi, L. 2007. "A Praxical Solution of the Symbol Grounding Problem." *Minds & Machines* 17: 369–389.

Tangiuchi, T. et al. 2016. "Symbol emergence in robotics: a survey." *Advanced Robotics* 30 (11–12): 706–728.

Tangiuchi, T. et al. 2019. "Survey on frontiers of language and robotics." *Advanced Robotics* 33 (6): 2–31.

Varela, F. J., Thompson, E. and Rosch, E. 1991. *The Embodied Mind: Cognitive Science and Human Experience.* Cambridge: The MIT Press.

Vogt, P. 2007. "Language Evolution and Robotics, Issues on Symbol Grounding and Language Acquisition." In A. Loula et al. (eds.). *Artificial Cognition Systems.* Hershey: Idea Group Publishing, 176–209.

Ziemke, T. 1999. "Rethinking Grounding." In A. Riegler et al. (eds.). *Understanding Representation in the Cognitive Sciences.* New York: Plenum Press, 177–180.